Network Formation with Local Complements and Global Substitutes: The Case of R&D Networks.

Konig, M.D.; Hsieh, Chih-Sheng; Liu, Xiaodong

2018

Supplement to "Network Formation with Local Complements and Global Substitutes: The Case of R&D Networks"

Chih-Sheng Hsieh[a], Michael D. König[b,c,d], Xiaodong Liu[e]

[a] *Department of Economics, Chinese University of Hong Kong, CUHK Shatin, Hong Kong, China.*

[b] *Centre for Economic Policy Research (CEPR), London, United Kingdom.*

[c] *ETH Zurich, Swiss Economic Institute (KOF), Zurich, Switzerland.*

[d] *Department of Spatial Economics, VU Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.*

[e] *Department of Economics, University of Colorado Boulder, Boulder, Colorado 80309–0256, United States.*

Contents

B. Definitions and Characterizations

A *network* (*graph*) G is the pair $(\mathcal{N}, \mathcal{E})$ consisting of a set of *nodes* (*vertices*) $\mathcal{N} = \{1, \ldots, n\}$ and a set of edges (*links*) $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ between them. A link (i, j) is *incident* with nodes i and j. The *neighborhood* of a node $i \in \mathcal{N}$ is the set $\mathcal{N}_i = \{j \in \mathcal{N} : (i, j) \in \mathcal{E}\}$. The *degree* d_i of a node $i \in \mathcal{N}$ gives the number of links incident to node i. Clearly, $d_i = |\mathcal{N}_i|$. Let $\mathcal{N}_i^{(2)} = \bigcup_{j \in \mathcal{N}_i} \mathcal{N}_j \setminus (\mathcal{N}_i \cup \{i\})$ denote the second-order neighbors of node i. Similarly, the k-th order neighborhood of node i is defined recursively from $\mathcal{N}_i^{(0)} = \{i\}$, $\mathcal{N}_i^{(1)} = \mathcal{N}_i$ and $\mathcal{N}_i^{(k)} = \bigcup_{j \in \mathcal{N}_i^{(k-1)}} \mathcal{N}_j \setminus \left(\bigcup_{l=0}^{k-1} \mathcal{N}_i^{(l)} \right)$. A *walk* in G of length k from i to j is a sequence $\langle i_0, i_1, \ldots, i_k \rangle$ of nodes such that $i_0 = i$, $i_k = j$, $i_p \neq i_{p+1}$, and i_p and i_{p+1} are (directly) linked, that is $i_p i_{p+1} \in \mathcal{E}$, for all $0 \leq p \leq k-1$. Nodes i and j are said to be *indirectly linked* in G if there exists a walk from i to j in G containing nodes other than i and j. A pair of nodes i and j is connected if they are either directly or indirectly linked. A node $i \in \mathcal{N}$ is *isolated* in G if $\mathcal{N}_i = \emptyset$. The network G is said to be *empty* (denoted by $\overline{K}_n$) when all its nodes are isolated.

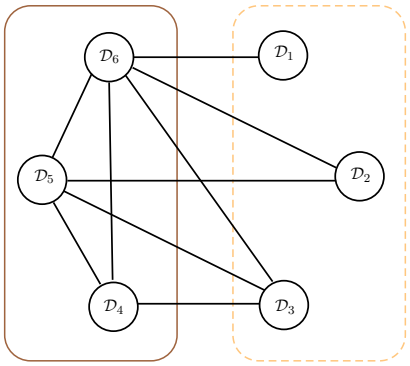
A *subgraph*, G', of G is the graph of subsets of the nodes, $\mathcal{N}(G') \subseteq \mathcal{N}(G)$, and links, $\mathcal{E}(G') \subseteq \mathcal{E}(G)$. A graph G is *connected*, if there is a path connecting every pair of nodes. Otherwise G is disconnected. The *components* of a graph G are the maximally connected subgraphs. A component is said to be *minimally connected* if the removal of any link makes the component disconnected.

Given a graph G and a set $\mathcal{S} \subseteq \mathcal{N}$, we say that $G_{\mathcal{S}}$ is the *subgraph* G *induced* S whenever the adjacency matrix of $G_{\mathcal{S}}$ is $\mathbf{A}_{\mathcal{S}}$. We write $G_{-\mathcal{S}}$ to denote the network $G_{\mathcal{N} \setminus \mathcal{S}}$, that is $G_{-\mathcal{S}}$ is the network that results after eliminating all the nodes in $\mathcal{S}$.

A *dominating set* for a graph $G = (\mathcal{N}, \mathcal{E})$ is a subset $\mathcal{S} \subseteq \mathcal{N}$ such that every node not in $\mathcal{S}$ is connected to at least one member of S by a link. An *independent set* is a set of nodes in a graph in which no two nodes are adjacent. For example the central node in a star $K_{1,n-1}$ forms a dominating set while the peripheral nodes form an independent set.

Let $G = (\mathcal{N}, \mathcal{E})$ be a graph whose distinct positive degrees are $d_{(1)} < d_{(2)} < \ldots < d_{(k)}$, and let $d_0 = 0$ (even if no agent with degree 0 exists in G). Further, define $\mathcal{D}_i = \{v \in \mathcal{N} : d_v = d_{(i)}\}$ for $i = 0, \ldots, k$. Then the set-valued vector $\mathcal{D} = (\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_k)$ is called the *degree partition* of G. A *nested split graph* is a graph with a nested neighborhood structure such that the set of neighbors of each node is contained in the set of neighbors of each higher degree node [Cvetkovic and Rowlinson, 1990; Mahadev and Peled, 1995]. Let $\mathcal{D} = (\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_k)$ be the degree partition of a nested split graph $G = (\mathcal{N}, \mathcal{E})$. Then the nodes $\mathcal{N}$ can be partitioned in independent sets $\mathcal{D}_i$, $i = 1, \ldots, \lfloor \frac{k}{2} \rfloor$ and a dominating set $\bigcup_{i=\lfloor \frac{k}{2} \rfloor + 1}^{k} \mathcal{D}_i$ in the graph $G' = (\mathcal{N} \setminus \mathcal{D}_0, \mathcal{E})$. Moreover, the neighborhoods of the nodes are nested. In particular, for each node $v \in \mathcal{D}_i$, $\mathcal{N}_v = \bigcup_{j=1}^{i} \mathcal{D}_{k+1-j}$ if $i = 1, \ldots, \lfloor \frac{k}{2} \rfloor$ if $i = 1, \ldots, k$, while $\mathcal{N}_v = \bigcup_{j=1}^{i} \mathcal{D}_{k+1-j} \setminus \{v\}$ if $i = \lfloor \frac{k}{2} \rfloor + 1, \ldots, k$. See also the left panel of Figure B.1.

In a *complete* graph K_n, every node is adjacent to every other node. The graph in which no pair of nodes is adjacent is the empty graph $\overline{K}_n$. A *clique* $K_{n'}$, $n' \leq n$, is a complete subgraph of the network G. A graph is *k-regular* if every node i has the same number of links $d_i = k$ for all $i \in \mathcal{N}$. The complete graph K_n is $(n-1)$-regular. The cycle C_n is 2-regular. In a *bipartite graph* there exists a partition of the nodes in two disjoint sets $\mathcal{V}_1$ and $\mathcal{V}_2$ such that each link connects a node in $\mathcal{V}_1$ to a node in $\mathcal{V}_2$. $\mathcal{V}_1$ and $\mathcal{V}_2$ are independent sets with cardinalities n_1 and n_2, respectively. In a complete bipartite graph K_{n_1,n_2} each node in $\mathcal{V}_1$ is connected to each other



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure B.1: (Left panel) Representation of nested split graphs and their degree partitions $\mathcal{D}$ with corresponding adjacency matrices $\mathbf{A}$. A line between $\mathcal{D}_i$ and $\mathcal{D}_j$ indicates that every node in $\mathcal{D}_i$ is adjacent to every node in $\mathcal{D}_j$. The partitions included in the solid frame ($\mathcal{D}_i$ with $\lfloor \frac{k}{2} \rfloor + 1 \leq i \leq k$) are the dominating subsets while the partitions in the dashed frame ($\mathcal{D}_i$ with $1 \leq i \leq \lfloor \frac{k}{2} \rfloor$) are the independent sets with $k = 6$. The illustration follows Mahadev and Peled [1995, p. 11]. (Right panel) The corresponding stepwise adjacency matrix $\mathbf{A}$ with elements a_{ij} satisfying the following condition: if $i < j$ and $a_{ij} = 1$ then $a_{hk} = 1$ whenever $h < k \leq j$ and $h \leq i$.

node in $\mathcal{V}_2$. The *star* $K_{1,n-1}$ is a complete bipartite graph in which $n_1 = 1$ and $n_2 = n - 1$.

The *complement* of a graph G is a graph $\overline{G}$ with the same nodes as G such that any two nodes of $\overline{G}$ are adjacent if and only if they are not adjacent in G. For example the complement of the complete graph K_n is the empty graph $\overline{K}_n$.

Let $\mathbf{A}$ be the symmetric $n \times n$ *adjacency matrix* of the network G. The element $a_{ij} \in \{0, 1\}$ indicates if there exists a link between nodes i and j such that $a_{ij} = 1$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ if $(i, j) \notin \mathcal{E}$. The k-th power of the adjacency matrix is related to walks of length k in the graph. In particular, $\left(\mathbf{A}^k\right)_{ij}$ gives the number of walks of length k from node i to node j. The *eigenvalues* of the adjacency matrix $\mathbf{A}$ are the numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ has a nonzero solution vector $\mathbf{v}_i$, which is an *eigenvector* associated with λ_i for $i = 1, \ldots, n$. Since the adjacency matrix $\mathbf{A}$ of an undirected graph G is real and symmetric, the eigenvalues of $\mathbf{A}$ are real, $\lambda_i \in \mathbb{R}$ for all $i = 1, \ldots, n$. Moreover, if $\mathbf{v}_i$ and $\mathbf{v}_j$ are eigenvectors for different eigenvalues, $\lambda_i \neq \lambda_j$, then $\mathbf{v_i}$ and $\mathbf{v}_j$ are orthogonal, i.e. $\mathbf{v}_i^\top \mathbf{v}_j = 0$ if $i \neq j$. In particular, $\mathbb{R}^n$ has an orthonormal basis consisting of eigenvectors of $\mathbf{A}$. Since $\mathbf{A}$ is a real symmetric matrix, there exists an orthogonal matrix $\mathbf{S}$ such that $\mathbf{S}^\top \mathbf{S} = \mathbf{S}\mathbf{S}^\top = \mathbf{I}$ (that is $\mathbf{S}^\top = \mathbf{S}^{-1}$) and $\mathbf{S}^\top \mathbf{A}\mathbf{S} = \mathbf{D}$, where $\mathbf{D}$ is the diagonal matrix of eigenvalues of $\mathbf{A}$ and the columns of $\mathbf{S}$ are the corresponding eigenvectors. The *Perron-Frobenius eigenvalue* $\lambda_{\mathrm{PF}}(G)$ is the *largest real eigenvalue* of $\mathbf{A}$ associated with G, i.e. all eigenvalues λ_i of $\mathbf{A}$ satisfy $|\lambda_i| \leq \lambda_{\mathrm{PF}}(G)$ for $i = 1, \ldots, n$ and there exists an associated nonnegative eigenvector $\mathbf{v}_{\mathrm{PF}} \geq 0$ such that $\mathbf{A}\mathbf{v}_{\mathrm{PF}} = \lambda_{\mathrm{PF}}(G)\mathbf{v}_{\mathrm{PF}}$. For a connected graph G the adjacency matrix $\mathbf{A}$ has a unique largest real eigenvalue $\lambda_{\mathrm{PF}}(G)$ and a positive associated eigenvector $\mathbf{v}_{\mathrm{PF}} > 0$. There exists a relation between the number of walks in a graph and its eigenvalues. The number of closed walks of length k from a node i in G to herself is given by $\left(\mathbf{A}^k\right)_{ii}$ and the total number of closed walks of length k in G is $\mathrm{tr}\left(\mathbf{A}^k\right) = \sum_{i=1}^n \left(\mathbf{A}^k\right)_{ii} = \sum_{i=1}^n \lambda_i^k$. We further have that $\mathrm{tr}\left(\mathbf{A}\right) = 0$, $\mathrm{tr}\left(\mathbf{A}^2\right)$ gives twice the number of links in G and $\mathrm{tr}\left(\mathbf{A}^3\right)$ gives six times the number of triangles in G.

A nested split graph is characterized by a *stepwise adjacency matrix* $\mathbf{A}$, which is a symmetric, binary $(n \times n)$-matrix with elements a_{ij} satisfying the following condition: if $i < j$ and $a_{ij} = 1$ then $a_{hk} = 1$ whenever $h < k \leq j$ and $h \leq i$. See also the right panel in Figure B.1. Both, the complete graph, K_n, as well as the star $K_{1,n-1}$, are particular examples of nested split graphs. Nested split graphs are also the graphs which maximize the largest eigenvalue, $\lambda_{\mathrm{PF}}(G)$, [Brualdi

and Solheid, 1986], and they are the ones that maximize the degree variance [Peled et al., 1999]. See e.g., König et al. [2014a] for further properties.

Given n and $p \in [0,1]$, the *random graph* $G(n,p)$ is generated by letting each pair of nodes be connected by an edge with probability p, independently. A natural generalization of $G(n,p)$ is obtained by replacing the single parameter p by a symmetric $n \times n$ matrix $(p_{ij})_{1 \le i,j,n}$ with $0 \le p_{ij} \le 1$. We write $G(n,(p_{ij}))$ for the *inhomogeneous random graph* with nodes set $\mathcal{N}$ where i and j are connected by a link with probability p_{ij}, and these events are independent for all pairs (i,j) with $1 \le i < j \le n$ [Bollobás et al., 2007; Söderberg, 2002].

Given a set of attributes $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, a vertex v is associated with the set $\mathcal{S}(v)$ of attributes selected by v from $\mathcal{A}$. Let the vertices $v_1, \ldots, v_n$ choose their attribute sets $\mathcal{S}_i = \mathcal{S}(v_i)$, $1 \le i \le n$, independently at random, and make v_i and v_j adjacent whenever they have at least $s \ge 1$ attributes in common, that is, $|\mathcal{S}_i \cap \mathcal{S}_j| \ge s$. The graph on the vertex set $\mathcal{N} = \{v_1, \ldots, v_n\}$ defined by this adjacency relationship is then called the *random intersection graph* $G(n,m,s)$ [Deijfen and Kets, 2009].

C. Cournot Competition and Profits from R&D Collaborations

We consider a Cournot oligopoly game in which a set $\mathcal{N} = \{1, \ldots, n\}$ of firms is competing in a homogeneous product market.[1] We assume that firms are not only competitors in the product market, but they can also form pairwise collaborative agreements.[2] These pairwise links involve a commitment to share R&D results and thus lead to lower marginal cost of production of the collaborating firms.[3] The amount of this cost reduction depends on the effort the firms invest into R&D. Given the collaboration network $G \in \mathcal{G}^n$, where $\mathcal{G}^n$ denotes the set of all graphs with n nodes, each firm sets an R&D effort level unilaterally.[4] Given the effort levels $e_i \ge 0$, marginal cost $c_i \ge 0$ of firm i is given by [Spence, 1984][5,6]

$$c_i(\mathbf{e}, G) = \bar{c}_i - \alpha e_i - \beta \sum_{j=1}^{n} a_{ij} e_j, \tag{C.1}$$

where $a_{ij} = 1$ if firms i and j set up a collaboration (0 otherwise) and $a_{ii} = 0$. The parameter

[1] Generalizations to Bertrand competition are straight forward [see König et al., 2018; Westbrock, 2010].

[2] Such R&D collaborations often involve competing firms, as for example a strategic alliance between *Pfizer* and *Bayer*, both operating in the pharmaceuticals sector (with primary standard industry classification code 2834) to develop treatments for obesity, type 2 diabetes and other related disorders in the year 2006 illustrates.

[3] R&D partnerships have become a widespread phenomenon characterizing technological dynamics, especially in industries with rapid technological development such as, for instance, the pharmaceutical and chemical industries [see e.g. Hagedoorn, 2002; Powell et al., 2005; Roijakkers and Hagedoorn, 2006]. In these industries firms have become more specialized in specific domains of a technology and they tend to combine their knowledge with that of other firms that are specialized in different domains in order to jointly generate innovations that can help to develop new products or reduce their production costs [Ahuja, 2000; Powell et al., 1996]. For example, Bernstein [1988] finds that R&D spillovers decrease the unit costs of production for a sample of Canadian firms. Similarly Belderbos et al. [2004] find evidence for production cost reductions due to R&D collaborations using data on a large sample of Dutch innovating firms.

[4] See also Kamien et al. [1992] for a similar model of competitive RJVs in which firms unilaterally choose their R&D effort levels.

[5] Note that we have neglected spillovers among non-collaborating firms. For an extension incorporating this additional spillover channel see König et al. [2018].

[6] This generalizes earlier studies such as the one by D'Aspremont and Jacquemin [1988] where spillovers were assumed to take place between all firms in the industry and no distinction between collaborating and non-collaborating firms was made.

$\alpha \geq 0$ measures the relative cost reduction due to a firms' own R&D effort while the parameter $\beta \geq 0$ measures the relative cost reduction due to the R&D effort of its collaboration partners.[7] We further allow for ex ante heterogeneity among firms in the variable cost $\bar{c}_i \geq 0$, for $i = 1, \ldots, n$ [see also Banerjee and Duflo, 2005], expressing their different technological and organizational capabilities.[8]

Moreover, we also assume that firms incur a direct cost γe_i^2, $\gamma \geq 0$, for their R&D efforts and a fixed cost $\zeta \geq 0$ for each R&D collaboration.[9] The profit of firm i, given the R&D network G and the quantities $0 \leq q_i \leq \bar{q}$ and efforts $0 \leq e_i \leq \bar{e}$, is then given by

$$\pi_i(\mathbf{q}, \mathbf{e}, G) = (p_i - c_i)q_i - \gamma e_i^2 - \zeta d_i, \tag{C.2}$$

with vectors $\mathbf{q} = (q_i)_{i=1}^n$ and $\mathbf{e} = (e_i)_{i=1}^n$. Inserting marginal costs from Equation (C.1) gives

$$\pi_i(\mathbf{q}, \mathbf{e}, G) = p_i q_i - \bar{c}_i q_i + \alpha q_i e_i + \beta q_i \sum_{j=1}^n a_{ij} e_j - \gamma e_i^2 - \zeta d_i.$$

The first-order condition with respect to R&D effort e_i is given by $\frac{\partial \pi_i(\mathbf{q},\mathbf{e},G)}{\partial e_i} = \alpha q_i - 2\gamma e_i = 0$. Solving for e_i and taking into account that $e_i \in [0, \bar{e}]$ delivers $e_i^* = \min\{\lambda q_i, \bar{e}\}$, where we have denoted by $\lambda = \frac{\alpha}{2\gamma}$. [10] This equation can be viewed as reflecting learning-by-doing effects on R&D efforts. It reflects various empirical studies which have found that the R&D effort of a firm is correlated with its output or size [Cohen and Klepper, 1996a,b]. We then can write marginal costs from Equation (C.1) as follows[11]

$$c_i(\mathbf{e}^*(\mathbf{q}), G) = \bar{c} - \lambda\alpha q_i - \lambda\beta \sum_{j=1}^n a_{ij} q_j. \tag{C.3}$$

Profits can be written as

$$\pi_i(\mathbf{q}, G) \equiv \pi_i(\mathbf{q}, \mathbf{e}^*(\mathbf{q}), G) = p_i q_i - \bar{c} q_i - \lambda\alpha q_i^2 + \lambda\beta q_i \sum_{j=1}^n a_{ij} q_j - \lambda^2 \gamma q_i^2 - \zeta d_i. \tag{C.4}$$

Next we consider the demand for goods produced by firm i. A representative consumer maximizes [Singh and Vives, 1984]

$$U(I, q_1, \ldots, q_n) = I + a \sum_{i=1}^n q_i - \frac{1}{2} \sum_{i=1}^n q_i^2 - \frac{b}{2} \sum_{i=1}^n \sum_{j \neq i}^n q_i q_j, \tag{C.5}$$

[7] Note that in this model, firms are exposed to business stealing effects if their rivals increase their output via cost reducing R&D collaborations.

[8] Blundell et al. [1995] argued that because the main source of unobserved heterogeneity in models of innovation lies in the different knowledge stocks with which firms enter a sample, a variable that approximates the build-up of firm knowledge at the time of entering the sample is a particularly good control for unobserved heterogeneity.

[9] In Section 2.4 we discuss several extensions of the model including heterogeneous linking costs.

[10] König et al. [2018] show that with $q_i \in [0, \bar{q}]$ we must have that $0 \leq e_i \leq q_i \leq \bar{q}$, and requiring that $\min_{i \in \mathcal{N}} \bar{c}_i > \bar{q}(1 + \beta(n-1))$, implies that the best response effort level of firm i is given by $e_i^* = \lambda q_i$.

[11] We assume that firms always implement the optimal R&D effort level. Since the optimal R&D effort decision only depends on a firm's own output, we assume that a firm does not face any uncertainty when implementing this strategy. In Section 2.2 we will, however, introduce noise in the optimal output and collaboration decisions, since these depend on the decisions of all other firms in the industry and their characteristics, which might be harder to observe.

with the budget constraint $I + \sum_{i=1}^{n} p_i q_i \leq E$, endowment E and a numeraire good I. The parameter a captures the total size of the market, whereas $b \in (0, 1]$, measures the degree of substitutability between products. In particular, $b = 1$ depicts a market of perfect substitutable goods, while $b \to 0$ represents the case of almost independent markets.[12] The constraint is binding and the utility maximization of the representative consumer gives the inverse demand function for firm i:[13]

$$p_i = a - q_i - b \sum_{j \neq i}^{n} q_j. \tag{C.6}$$

Firm i then sets its quantity, q_i, in order to maximize its profit, π_i, given by Equation (C.4). We also assume that there is a maximum production capacity $\overline{q}$ such that $q_i \leq \overline{q}$ for all $i \in \mathcal{N}$. Inserting marginal cost from Equation (C.3) and inverse demand from Equation (C.6) we can write firm i's profit as

$$\pi_i(\mathbf{q}, G) = (a - \bar{c}_i)q_i - (1 - \lambda\alpha + \lambda^2\gamma)q_i^2 - bq_i \sum_{j \neq i} q_j + \lambda\beta \sum_{j=1}^{n} a_{ij} q_i q_j - \zeta d_i. \tag{C.7}$$

In the following we will denote by $\eta_i \equiv a - \bar{c}_i$, $\nu \equiv 1 - \lambda\alpha + \lambda^2\gamma$ and $\rho \equiv \lambda\beta$, so that Equation (C.7) becomes [Ballester et al., 2006]

$$\pi_i(\mathbf{q}, G) = \underbrace{\eta_i q_i - \nu q_i^2}_{\text{own concavity}} \quad \underbrace{-bq_i \sum_{j \neq i}^{n} q_j}_{\text{global substitutability}} \quad + \quad \underbrace{\rho q_i \sum_{j=1}^{n} a_{ij} q_j}_{\text{local complementarity}} \quad -\zeta d_i,$$

which is exactly Equation (1) in Section 2.

D. Multinomial Logit Output Choice

Consider a discretization (partition) $\mathcal{Q}_N = \{0, \Delta q, 2\Delta q, \ldots, \overline{q}\}$ of the interval $\mathcal{Q} = [0, \overline{q}]$ into N subintervals with length $\Delta q = \overline{q}/N$ for some (large) N. Let the profit of firm i from choosing an output level $q_i \in \mathcal{Q}_N$ be given by $\pi_i(q_i, \mathbf{q}_{-i}, G) + \varepsilon_i$. When the error term ε_i is independently and identically type-I extreme value distributed with parameter ϑ we get [Anderson et al., 1992]:[14]

$$\mathbb{P}\left(q_i = \underset{q_i' \in \mathcal{Q}_s}{\operatorname{argmax}} \left\{ \pi_i(q_i, \mathbf{q}_{-i}, G) + \varepsilon_i \right\} \right) = \frac{e^{\vartheta \pi_i(q_i, \mathbf{q}_{-i}, G)}}{\sum_{q_i' \in \mathcal{Q}_N} e^{\vartheta \pi_i(q_i', \mathbf{q}_{-i}, G)}}.$$

[12] Observe that we do not ex ante impose any restrictions on the parameter b, in particular, we do not require that $b = 1$ nor $b = 0$.

[13] With the budget constraint $I = E - \sum_{i=1}^{n} p_i q_i$ in the consumer's utility in Equation (C.5), the FOC is given by $\frac{\partial U}{\partial q_i} = -p_i + a - q_i - b\sum_{j \neq i}^{n} q_j = 0$, from which we obtain Equation (C.6). Moreover, inserting into the utility of the consumer gives: $U = E - \sum_{i=1}^{n} p_i q_i + a\sum_{i=1}^{n} q_i - \frac{1}{2}\sum_{i=1}^{n} q_i^2 - \frac{b}{2}\sum_{i=1}^{n}\sum_{j \neq i}^{n} q_i q_j = E - \sum_{i=1}^{n}\left(a - q_i - b\sum_{j \neq i}^{n} q_j\right) q_i + a\sum_{i=1}^{n} q_i - \frac{1}{2}\sum_{i=1}^{n} q_i^2 - \frac{b}{2}\sum_{i=1}^{n}\sum_{j \neq i}^{n} q_i q_j = E + \frac{1}{2}\sum_{i=1}^{n} q_i^2 + \frac{b}{2}\sum_{i=1}^{n}\sum_{j \neq i}^{n} q_i q_j$. This defines the consumer surplus that has been introduced in Section 2.3.

[14] For a computationally efficient implementation of the multinomial logit model see D'Haultfœuille and Iaria [2016].

Assume that the output adjustment rate is given by $\chi N/\bar{q} > 0$. Then

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{\omega}_{t+\Delta t} = (q_i, \mathbf{q}_{-it}, G_t) | \boldsymbol{\omega}_t = (\mathbf{q}_t, G_t)\right) &= \frac{\chi N}{\bar{q}} \mathbb{P}\left(q_i = \underset{q_i' \in \mathcal{Q}_s}{\operatorname{argmax}} \left\{\pi_i(q_i', \mathbf{q}_{-i}, G) + \varepsilon\right\}\right) \Delta t + o(\Delta t) \\ &= \frac{\chi N}{\bar{q}} \frac{e^{\vartheta \pi_i(q_i, \mathbf{q}_{-it}, G_t)}}{\sum_{q_i' \in \mathcal{Q}_N} e^{\vartheta \pi_i(q_i', \mathbf{q}_{-it}, G_t)}} \Delta t + o(\Delta t) \\ &= \chi \frac{e^{\vartheta \pi_i(q_i, \mathbf{q}_{-it}, G_t)}}{\sum_{q_i' \in \mathcal{Q}_N} e^{\vartheta \pi_i(q_i', \mathbf{q}_{-it}, G_t)} \Delta q} \Delta t + o(\Delta t). \end{aligned} \tag{D.8}$$

Using the trapezoidal rule we can write the sum in the denominator in the last line of Equation (D.8) as follows [see e.g., Atkinson, 1989]

$$\sum_{q_i' \in \mathcal{Q}_N} e^{\vartheta \pi_i(q_i', \mathbf{q}_{-it}, G_t)} \Delta q = \int_{\mathcal{Q}} e^{\vartheta \pi_i(q', \mathbf{q}_{-it}, G_t)} dq' + \frac{\Delta q}{2} \left(e^{\vartheta \pi_i(0, \mathbf{q}_{-it}, G_t)} + e^{\vartheta \pi_i(\bar{q}, \mathbf{q}_{-it}, G_t)}\right) + o(\Delta q). \tag{D.9}$$

Then in the limit of $N \to \infty$, respectively, $\Delta q \downarrow 0$, we can write

$$\mathbb{P}\left(\boldsymbol{\omega}_{t+\Delta t} = (q_i, \mathbf{q}_{-it}, G_t) | \boldsymbol{\omega}_t = (\mathbf{q}_t, G_t)\right) = \chi \frac{e^{\vartheta \pi_i(q, \mathbf{q}_{-it}, G_t)}}{\int_{\mathcal{Q}} e^{\vartheta \pi_i(q', \mathbf{q}_{-it}, G_t)} dq'} \Delta t + o(\Delta t),$$

which is exactly Equation (3). See also Anderson et al. [2004, 2001]; Ben-Akiva and Watanatada [1981]; McFadden [1976] for further discussion. The transition probability then states that with increasing values of ϑ (lower levels of noise), firms choose output levels with higher probability that yield higher profits.

E. Firm Heterogeneity

In the following sections we will discuss three possible extensions of the model that incorporate firm heterogeneity. First, in Section E.1 we allow for heterogeneous collaborations costs. Second, in Section E.2 we incorporate heterogenous spillovers between collaborating firms.

E.1. Heterogeneous Marginal Collaboration Costs

In the following we assume that the marginal cost of collaboration between firms i and j can be written as $\zeta_i(G) = \sum_{j=1}^n a_{ij}\psi_{ij}$ (see Equation (23)), where the function ψ_{ij} is additively separable $\psi_{ij} = s_i + s_j$, and we assume that the cost $s_i \geq 0$ is proportional to the inverse of the firm's productivity, that is $s_i = \frac{1}{\phi_i}$, where $\phi_i > 0$ is the productivity (or efficiency) of firm i [similar to e.g., Melitz, 2003; Melitz et al., 2008]. This implies that firms with higher productivity incur lower collaboration costs. The probability of a link between firms i and j is then given by Equation (16):[15]

$$p^{\vartheta}(q_i, s_i, q_j, s_j) = \frac{e^{\vartheta(\rho q_i q_j - s_i - s_j)}}{1 + e^{\vartheta(\rho q_i q_j - s_i - s_j)}}. \tag{E.10}$$

[15] Similar specifications can be found in the empirical literature on network formation [Graham, 2015]. For example, Graham [2014] and Fafchamps and Gubert [2007] consider an econometric network formation model in which the probability of a link between agents i and j is given by $\mathbb{P}(a_{ij} = 1) = \frac{e^{X_i + X_j + \mathbf{z}_{ij}^\top \boldsymbol{\beta}}}{1 + e^{X_i + X_j + \mathbf{z}_{ij}^\top \boldsymbol{\beta}}}$ where X_i is an agent specific fixed effect and $\mathbf{Z}_{ij}$ is a vector of pair specific covariates. Similarly, Chatterjee et al. [2011] analyze a network formation model with linking probability $\mathbb{P}(a_{ij} = 1) = \frac{e^{X_i + X_j}}{1 + e^{X_i + X_j}}$.

Next, we assume that the firms' productivities, $\phi_i \geq c > 0$, are Pareto distributed [König et al., 2016; Melitz et al., 2008], with density $f(x) = \frac{\gamma}{c}\left(\frac{c}{x}\right)^{\gamma+1}$ with $x > c$, where $c > 0$ is a lower-cut-off and $\gamma > 0$ is a positive parameter. The complementary distribution function is then given by $F(x) = 1 - \left(\frac{c}{x}\right)^{\gamma}$. It follows that the cost $s = \frac{1}{\phi}$ has the density $f(s) = \gamma c^{\gamma} s^{\gamma-1}$ for $s \in \left(0, \frac{1}{c}\right)$, and the cumulative distribution function $F(s) = (cs)^{\gamma}$.[16]

Proposition 5. *Assume that the firms output levels are concentrated on q^* in the limit of $\vartheta \to \infty$, then under the continuum approximation,[17] the degree distribution is given by*

$$P(k) = \frac{c^{\gamma}}{k}\left(\rho(q^*)^2 - \left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)^{\gamma-1}\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}, \tag{E.11}$$

and for large k the degree distribution $P(k)$ decays as $O\left(k^{-\frac{\gamma-1}{\gamma}}\right)$.

Proof of Proposition 5. The generating function of the degree $d_1(G)$ is given by

$$\begin{aligned}\mathbb{E}\left(x_1^{d_1(G)}\right) &= \mathbb{E}\left(\mathbb{E}\left(x_1^{d_1(G)}\middle| q_1, s_1\right)\right)\\ &= \mathbb{E}\left(\left(\mathbb{E}\left(\frac{1+p(q_1,s_1,q_2,s_2)x_1}{1+p(q_1,s_1,q_2,s_2)}\middle| q_1, s_1\right)\right)^{n-1}\right),\end{aligned}$$

With the cost distributed as $f(s) = \gamma c^{\gamma} s^{\gamma-1}$ for $s \in \left(0, \frac{1}{c}\right)$, we can write

$$\begin{aligned}\mathbb{E}\left(x_1^{d_1(G)}\middle| q_1 = q, s_1 = s\right) &= \mathbb{E}\left(\frac{1+p(q,s,q_2,s_2)x_1}{1+p(q,s,q_2,s_2)}\middle| q, s\right)^{n-1}\\ &= \left(1 + (x_1-1)\mathbb{E}\left(p(q,s,q_2,s_2)\middle| q_1 = q, s_1 = s\right)\right)^{n-1}\\ &= \left(1 + (x_1-1)\int_{\mathcal{Q}} dq'\mu^{\vartheta}(q')\int ds'\gamma c^{\gamma}(s')^{\gamma-1}p(q,s,q',s')\right)^{n-1}\\ &= \left(1 + (x_1-1)\int_{\mathcal{Q}} dq'\mu^{\vartheta}(q')\int ds'\gamma c^{\gamma}(s')^{\gamma-1}\frac{e^{\vartheta(\rho qq'-s-s')}}{1+e^{\vartheta(\rho qq'-s-s')}}\right)^{n-1}.\end{aligned}$$

In the limit of $\vartheta \to \infty$ in Equation (E.10) we obtain

$$\lim_{\vartheta\to\infty} p^{\vartheta}(q,s,q',s') = \lim_{\vartheta\to\infty}\frac{e^{\vartheta(\rho qq'-s-s')}}{1+e^{\vartheta(\rho qq'-s-s')}} = \mathbb{1}_{\{\rho qq'>s+s'\}},$$

so that we can write

$$\begin{aligned}\mathbb{E}\left(x_1^{d_1(G)}\middle| q_1 = q, s_1 = s\right) &= \left(1 + (x_1-1)\int_{\mathcal{Q}} dq'\mu^{\vartheta}(q')\int ds'\gamma c^{\gamma}(s')^{\gamma-1}\mathbb{1}_{\{\rho qq'>s+s'\}}\right)^{n-1}\\ &= \left(1 + (x_1-1)\gamma c^{\gamma}\int_0^{\rho qq^*-s} ds'(s')^{\gamma-1}\right)^{n-1}\\ &= \left(1 + (x_1-1)c^{\gamma}\left(\rho qq^*-s\right)^{\gamma}\right)^{n-1}\\ &= e^{(x_1-1)(n-1)c^{\gamma}(\rho qq^*-s)^{\gamma}}.\end{aligned}$$

[16] In the following propositions we will assume that that the firms' output levels are concentrated on q^* in the limit of $\vartheta \to \infty$, and assumption that is typically satisfied in the simulation studies that we did. Moreover, concentration can be shown to hold in the basic model with homogeneous firms that has been analyzed in the main part of the paper.

[17] This is an approximation which has shown to be accurate in various network formation models as the network size becomes large [Dorogovtsev and Mendes, 2013, pp. 117].

This is the generating function of a Poisson random variable with expectation and variance given by $\bar{d}(q,s) \equiv (n-1)c^{\gamma}\left(\rho q q^{*}-s\right)^{\gamma}$. When the cut-off c is small, the variance becomes small, and we can approximate the Poisson random variable with a constant random variable at the expected value. Making further a continuum approximation, where we treat the degree as a continuous variable, we can write

$$\mathbb{P}\left(d_1(G)=k|\, q_1=q, s_1=s\right)=\delta\left(k-\bar{d}(q,s)\right)=\delta\left(k-(n-1)c^{\gamma}\left(\rho q q^{*}-s\right)^{\gamma}\right).$$

Note that under the continuum approximation there exists a one-to-one mapping from the degree k to the cost s, where for a given k and output q, the cost s is given by

$$s=\rho q q^{*}-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}.$$

Using the fact that[18]

$$\delta\left(k-(n-1)c^{\gamma}\left(\rho(q^{*})^{2}-s\right)^{\gamma}\right)=\delta\left(s-\left(\rho(q^{*})^{2}-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)\right)\frac{1}{\gamma k}\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}. \tag{E.12}$$

and assuming that the output distribution concentrates on q^{*}, the degree distribution is given by

$$\begin{aligned}
P(k) &= \int ds\mathbb{P}\left(d_1(G)=k|\, q_1=q^{*}, s_1=s\right) f(s) \\
&= \gamma c^{\gamma}\int ds\delta\left(s-(n-1)c^{\gamma}\left(\rho(q^{*})^{2}-s\right)^{\gamma}\right)s^{\gamma-1} \\
&= \gamma c^{\gamma}\int ds\delta\left(k-\left(\rho(q^{*})^{2}-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)\right)\frac{1}{\gamma k}\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}s^{\gamma-1} \\
&= \frac{c^{\gamma}}{k}\left(\rho(q^{*})^{2}-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)^{\gamma-1}\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}} \\
&= O\left(k^{-\frac{\gamma-1}{\gamma}}\right).
\end{aligned}$$

□

Hence, we obtain a power law degree distribution with parameter $\frac{\gamma-1}{\gamma}$, consistent with previous empirical studies which have found power law degree distributions in R&D alliance networks [e.g., Powell et al., 2005]. An illustration can be seen in Figure E.1 for the case of $\gamma=2$ and $n=200$ firms.

Proposition 6. *Assume that the firms output levels are concentrated on q^{*} in the limit of $\vartheta\to\infty$, then under the continuum approximation, the averge nearest neighbor degree distribution is given by*

$$k_{\mathrm{nn}}(k)=1+\frac{(n-1)^{2}\gamma c^{2\gamma}}{k}\int_{0}^{\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}}ds'(s'^{\gamma-1}(\rho(q^{*})^{2}-s'^{\gamma}. \tag{E.13}$$

[18] When $g(x)$ is a continuously differentiable function in $\mathbb{R}$ it holds that $\delta(g(x))=\sum_{i=1}^{m}\frac{\delta(x-x_i)}{|g'(x_i)|}$ where the m roots x_i satisfy $g(x_i)=0$ for all $i=1,\ldots,m$.

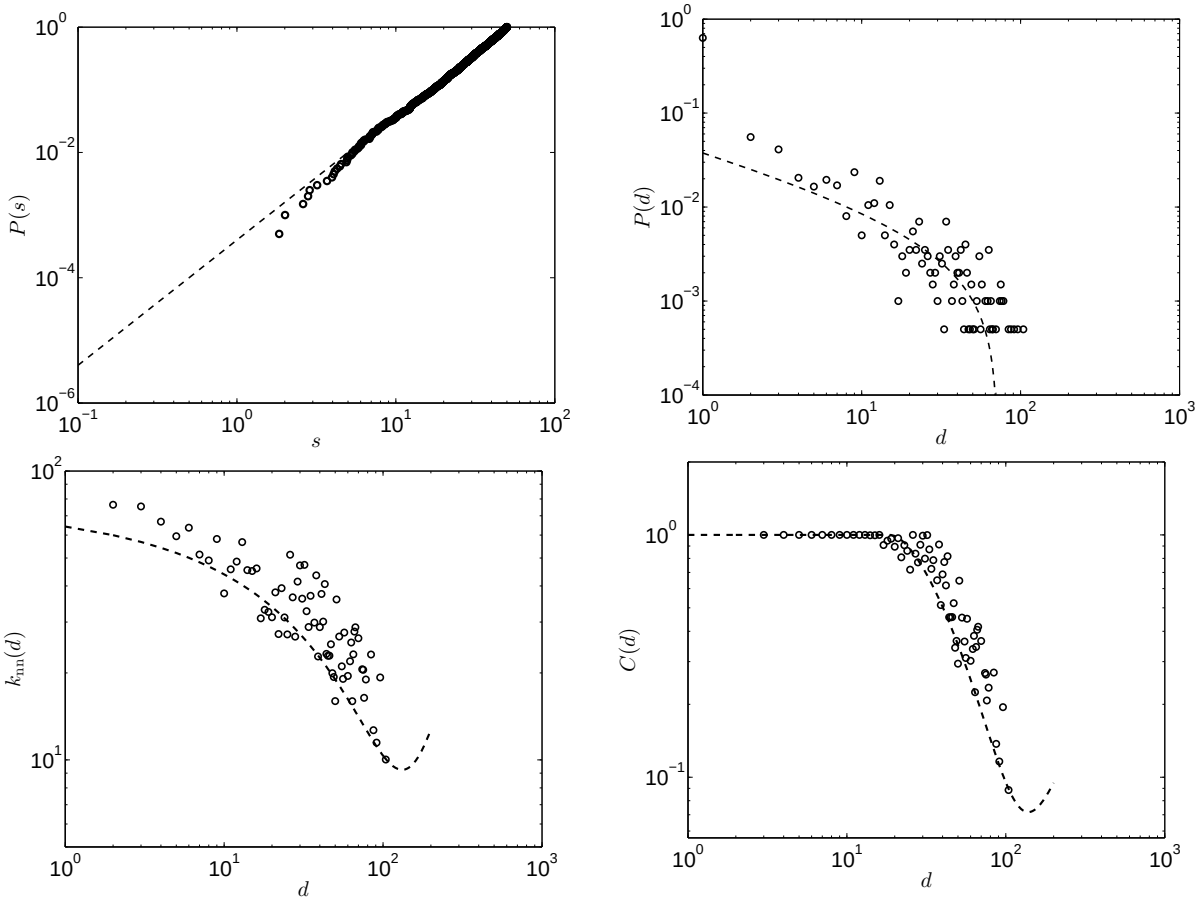


Figure E.1: (Top left panel) The empirical and the theoretical cumulative cost distribution $F(s) = (cs)^{\gamma}$ with $\gamma = 2$ and $c = 0.02$. The empirical distribution is indicated with circles and the theoretical distribution with a dashed line. (Top right panel) The degree distribution $P(d)$. The dashed line indicates the theoretical prediction of Equation (E.11). (Bottom left panel) The average nearest neighbor degree distribution $k_{\mathrm{nn}}(d)$, decreasing with increasing degrees d and thus indicating a dissortative network. The dashed line indicates the theoretical prediction of Equation (E.13). (Bottom right panel) The clustering degree distribution $C(d)$, decreasing with increasing degree d. The parameters used are $b = 0.75$, $\nu = 1$ and $\rho = 1$. The distributions are computed across 10 independent simulation runs with $n = 200$ firms.

Proof of Proposition 6. Next we compute the average nearest neighbor degree distribution

$$k_{\mathrm{nn}}(k) = 1 + \frac{1}{P(k)} \int ds \int_{\mathcal{Q}} dq f(s) \mu^{\vartheta}(q) g(k|q,s) \tilde{k}_{\mathrm{nn}}(q,s),$$

where

$$\begin{aligned}
\tilde{k}_{\mathrm{nn}}(q,s) &= \int ds' \int_{\mathcal{Q}} dq' p(q',s'|q,s) \bar{d}(q',s'), \\
g(k|q,s) &= \mathbb{P}\left(d_1(G) = k | \, q_1 = q, s_1 = s\right) = \delta\left(k - \bar{d}(q,s)\right), \\
\bar{d}(q,s) &= (n-1) c^{\gamma} \left(\rho q q^* - s\right)^{\gamma} \\
p(q',s'|q,s) &= \frac{(n-1) p(q,s,q',s') f(s') \mu^{\vartheta}(q')}{\bar{d}(q,s)}, \\
\lim_{\vartheta \to \infty} p^{\vartheta}(q,s,q',s') &= \lim_{\vartheta \to \infty} \frac{e^{\vartheta(\rho q q' - s - s')}}{1 + e^{\vartheta(\rho q q' - s - s')}} = \mathbb{1}_{\{\rho q q' > s + s'\}} \\
\mu^{\vartheta}(q) &= \delta(q - q^*) \\
f(s) &= \gamma c^{\gamma} s^{\gamma - 1}.
\end{aligned}$$

It then follows that

$$p(q',s'|q,s) = \frac{(n-1) \mathbb{1}_{\{\rho q q' > s + s'\}} \gamma c^{\gamma} (s')^{\gamma - 1} \delta(q' - q^*)}{\bar{d}(q,s)},$$

and therefore

$$\begin{aligned}
\tilde{k}_{\mathrm{nn}}(q,s) &= \frac{n-1}{\bar{d}(q,s)} \int ds' f(s') \int dq' \delta(q - q^*) \mathbb{1}_{\{\rho q q' > s + s'\}} \bar{d}(q',s') \\
&= \frac{n-1}{\bar{d}(q,s)} \int ds' f(s') \mathbb{1}_{\{\rho q q^* > s + s'\}} \bar{d}(q^*,s') \\
&= \frac{n-1}{\bar{d}(q,s)} \int_0^{\rho q q^* - s} ds' f(s') \bar{d}(q^*,s') \\
&= c^{-\gamma} (\rho q q^* - s)^{-\gamma} \int_0^{\rho q q^* - s} ds' \gamma c^{\gamma} (s')^{\gamma - 1} (n-1) c^{\gamma} (\rho (q^*)^2 - s')^{-\gamma} \\
&= \frac{(n-1) \gamma c^{\gamma} J(q,s)}{(\rho q q^* - s)^{\gamma}},
\end{aligned}$$

where we have denoted

$$J(q,s) \equiv \int_0^{\rho q q^* - s} ds' (\rho (q^*)^2 - s')^{\gamma} (s')^{\gamma - 1}.$$

We then get

$$\begin{aligned}
k_{\mathrm{nn}}(k) &= 1 + \frac{1}{P(k)} \int ds \int_{\mathcal{Q}} dq f(s) \delta(q - q^*) \delta\left(k - \bar{d}(q,s)\right) \tilde{k}_{\mathrm{nn}}(q,s) \\
&= 1 + \frac{1}{P(k)} \int ds \gamma c^{\gamma} s^{\gamma - 1} \delta\left(k - \bar{d}(q^*,s)\right) \tilde{k}_{\mathrm{nn}}(q^*,s) \\
&= 1 + \frac{1}{P(k)} \int ds f(s) \delta\left(k - \bar{d}(q^*,s)\right) \frac{(n-1) \gamma c^{\gamma} J(q^*,s)}{(\rho q q^* - s)^{\gamma}}.
\end{aligned}$$

Using Equation (E.12) we can write this as

$$\begin{aligned}
k_{\text{nn}}(k) &= 1 + \frac{1}{P(k)}\gamma c^{\gamma}\left(\rho(q^*)^2 - \left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)^{\gamma-1}\frac{1}{\gamma k}\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}} \\
&\times (n-1)c^{\gamma}\gamma J\left(q^*, \rho(q^*)^2 - \left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right) \\
&= 1 + \frac{(n-1)^2\gamma c^{2\gamma}}{k} J\left(q^*, \rho(q^*)^2 - \left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right) \\
&= 1 + \frac{(n-1)^2\gamma c^{2\gamma}}{k}\int_0^{\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}} ds'(s')^{\gamma-1}(\rho(q^*)^2 - s')^{\gamma}.
\end{aligned}$$

□

Figure E.1 shows the results from numerical simulations compared with the theoretical prediction of Equation (E.13).

Proposition 7. *Assume that the firms output levels are concentrated on q^* in the limit of $\vartheta \to \infty$, then under the continuum approximation, the clustering coefficient is given by*

$$\begin{aligned}
C(k) &= \mathbb{1}_{\left\{k<(n-1)\left(\frac{\rho(q^*)^2 c}{2}\right)^{\gamma}\right\}} + \mathbb{1}_{\left\{k>(n-1)\left(\frac{\rho(q^*)^2 c}{2}\right)^{\gamma}\right\}}\frac{(n-1)c^{\gamma}}{k} \\
&\times \left(1 + \gamma\frac{(n-1)c^{\gamma}}{k}\int_{\rho(q^*)^2 - \left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}}^{\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}} ds'(s'^{\gamma-1}\left(\rho(q^*)^2 - s'\right)^{\gamma}\right),
\end{aligned} \tag{E.14}$$

and for large k the clustering coefficient $C(k)$ decays as $O\left(\frac{1}{k}\right)$.

Proof of Proposition 7. Next we analyze the clustering coefficient of a firm with degree k, which can be written as

$$\begin{aligned}
C(k) &= \frac{1}{P(k)}\int ds \int_{\mathcal{Q}} dq f(s)\delta(q - q^*)g(k|q,s)\tilde{C}(q,s) \\
&= \frac{1}{P(k)}\int ds f(s)g(k|q^*,s)\tilde{C}(q^*,s),
\end{aligned}$$

where

$$\tilde{C}(q^*,s) = \int ds' \int ds'' \int_{\mathcal{Q}} dq' \int_{\mathcal{Q}} dq'' p(q',s',q'',s'')p(q',s'|q^*,s)p(q'',s''|q^*,s).$$

This can further be written as follows

$$\begin{aligned}
\tilde{C}(q^*,s) &= \int ds' \int ds'' \int_{\mathcal{Q}} dq' \int_{\mathcal{Q}} dq'' \mathbb{1}_{\{\rho q' q'' > s' + s''\}} \\
&\times \frac{(n-1)\mathbb{1}_{\{\rho q^* q' > s+s'\}} f(s')\delta(q' - q^*)}{\bar{d}(q^*,s)} \frac{(n-1)\mathbb{1}_{\{\rho q^* q'' > s+s''\}} f(s'')\delta(q'' - q^*)}{\bar{d}(q^*,s)} \\
&= \frac{(n-1)^2}{\bar{d}(q^*,s)^2}\int ds' f(s') \int ds'' f(s'')\mathbb{1}_{\{\rho(q^*)^2 > s'+s''\}}\mathbb{1}_{\{\rho(q^*)^2 > s+s'\}}\mathbb{1}_{\{\rho(q^*)^2 > s+s''\}} \\
&= \frac{(n-1)^2}{\bar{d}(q^*,s)^2}\left(\mathbb{1}_{\left\{s > \frac{\rho(q^*)^2}{2}\right\}}\int_0^{\rho(q^*)^2 - s} ds' f(s') \int_0^{\rho(q^*)^2 - s} ds'' f(s'')\right. \\
&\left. + \mathbb{1}_{\left\{s < \frac{\rho(q^*)^2}{2}\right\}}\left(\int_0^{s} ds' f(s') \int_0^{\rho(q^*)^2 - s} ds'' f(s'') + \int_s^{\rho(q^*)^2 - s} ds' f(s') \int_0^{\rho(q^*)^2 - s'} ds'' f(s'')\right)\right).
\end{aligned}$$

We then get (see also Figure E.2)

$$\begin{aligned}
\tilde{C}(q^*,s) &= \frac{\gamma^2 c^{2\gamma}(n-1)^2}{\bar{d}(q^*,s)}\left(\mathbb{1}_{\left\{s<\frac{\rho(q^*)^2}{2}\right\}}\int_0^{\rho(q^*)^2-s} ds'(s')^{\gamma-1}\int_0^{\rho(q^*)^2-s} ds''(s'')^{\gamma-1}\right.\\
&\left.+\mathbb{1}_{\left\{s>\frac{\rho(q^*)^2}{2}\right\}}\left(\int_0^{s} ds'(s')^{\gamma-1}\int_0^{\rho(q^*)^2-s} ds''(s'')^{\gamma-1}+\int_s^{\rho(q^*)^2-s} ds'(s')^{\gamma-1}\int_0^{\rho(q^*)^2-s'} ds''(s'')^{\gamma-1}\right)\right)\\
&= \frac{\gamma c^{2\gamma}(n-1)^2}{\bar{d}(q^*,s)}\left(\mathbb{1}_{\left\{s<\frac{\rho(q^*)^2}{2}\right\}}\int_0^{\rho(q^*)^2-s} ds'(s')^{\gamma-1}\left(\rho(q^*)^2-s\right)^{\gamma}\right.\\
&\left.+\mathbb{1}_{\left\{s>\frac{\rho(q^*)^2}{2}\right\}}\left(\int_0^{s} ds'(s')^{\gamma-1}\left(\rho(q^*)^2-s\right)^{\gamma}+\int_s^{\rho(q^*)^2-s} ds'(s')^{\gamma-1}\left(\rho(q^*)^2-s'\right)^{\gamma}\right)\right)\\
&= \frac{\gamma c^{2\gamma}(n-1)^2}{\bar{d}(q^*,s)}\left(\mathbb{1}_{\left\{s<\frac{\rho(q^*)^2}{2}\right\}}\frac{1}{\gamma}\left(\rho(q^*)^2-s\right)^{2\gamma}\right.\\
&\left.+\mathbb{1}_{\left\{s>\frac{\rho(q^*)^2}{2}\right\}}\left(\frac{1}{\gamma}s^{\gamma}\left(\rho(q^*)^2-s\right)^{\gamma}+\int_s^{\rho(q^*)^2-s} ds'(s')^{\gamma-1}\left(\rho(q^*)^2-s'\right)^{\gamma}\right)\right)\\
&= \frac{c^{2\gamma}(n-1)^2}{\bar{d}(q^*,s)}\left(\mathbb{1}_{\left\{s<\frac{\rho(q^*)^2}{2}\right\}}\left(\rho(q^*)^2-s\right)^{2\gamma}+\mathbb{1}_{\left\{s>\frac{\rho(q^*)^2}{2}\right\}}\left(s^{\gamma}\left(\rho(q^*)^2-s\right)^{\gamma}+\gamma J(s)\right)\right),
\end{aligned} \tag{E.15}$$

where we have denoted

$$J(s)\equiv\int_s^{\rho(q^*)^2-s} ds'(s')^{\gamma-1}\left(\rho(q^*)^2-s'\right)^{\gamma}.$$

Using the fact that $\bar{d}(q,s)\equiv(n-1)c^{\gamma}\left(\rho q q^*-s\right)^{\gamma}$ this can be written as

$$\tilde{C}(q^*,s)=\frac{1}{\left(\rho(q^*)^2-s\right)^{2\gamma}}\left(\mathbb{1}_{\left\{s<\frac{\rho(q^*)^2}{2}\right\}}\left(\rho(q^*)^2-s\right)^{2\gamma}+\mathbb{1}_{\left\{s>\frac{\rho(q^*)^2}{2}\right\}}\left(s^{\gamma}\left(\rho(q^*)^2-s\right)^{\gamma}+\gamma J(s)\right)\right). \tag{E.16}$$

Hence we get

$$\begin{aligned}
C(k) &= \frac{1}{P(k)}\int ds f(s)\delta\left(k-\bar{d}(q^*,s)\right)\tilde{C}(q^*,s)\\
&= \frac{1}{P(k)}\int ds\gamma c^{\gamma}s^{\gamma-1}\delta\left(s-\left(\rho(q^*)^2-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)\right)\frac{1}{\gamma k}\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\tilde{C}(q^*,s)\\
&= \frac{1}{P(k)}\gamma c^{\gamma}\left(\rho(q^*)^2-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)^{\gamma-1}\frac{1}{\gamma k}\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\tilde{C}\left(q^*,\rho(q^*)^2-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)\\
&= \tilde{C}\left(q^*,\rho(q^*)^2-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right).
\end{aligned}$$

Inserting Equation (E.16) this gives

$$\begin{aligned}
C(k) &= \mathbb{1}_{\left\{k<(n-1)\left(\frac{\rho(q^*)^2c}{2}\right)^{\gamma}\right\}}+\mathbb{1}_{\left\{k>(n-1)\left(\frac{\rho(q^*)^2c}{2}\right)^{\gamma}\right\}}\left(\frac{(n-1)c^{\gamma}}{k}+\gamma\left(\frac{(n-1)c^{\gamma}}{k}\right)^2 J\left(\rho(q^*)^2-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}\right)\right)\\
&= \mathbb{1}_{\left\{k<(n-1)\left(\frac{\rho(q^*)^2c}{2}\right)^{\gamma}\right\}}+\mathbb{1}_{\left\{k>(n-1)\left(\frac{\rho(q^*)^2c}{2}\right)^{\gamma}\right\}}\left(\frac{(n-1)c^{\gamma}}{k}\right.\\
&\left.+\gamma\left(\frac{(n-1)c^{\gamma}}{k}\right)^2\int_{\rho(q^*)^2-\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}}^{\left(\frac{k}{(n-1)c^{\gamma}}\right)^{\frac{1}{\gamma}}} ds'(s')^{\gamma-1}\left(\rho(q^*)^2-s'\right)^{\gamma}\right)
\end{aligned}$$

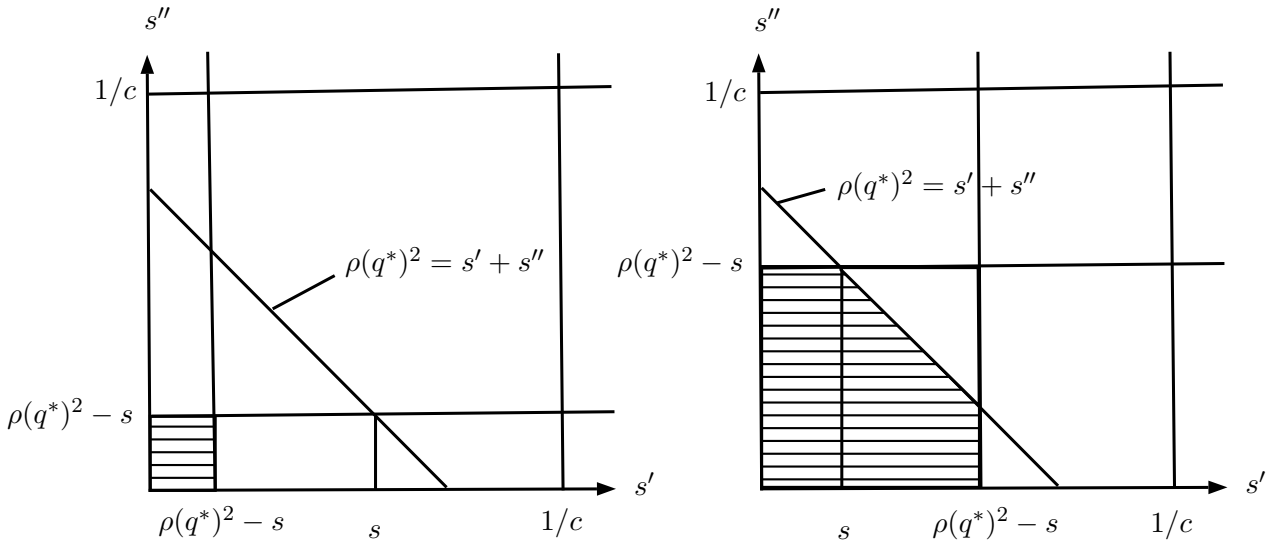


Figure E.2: The area of integration in Equation (E.15) for the case of $s < \frac{\rho(q^*)^2}{2}$ (left panel) and for the case of $s > \frac{\rho(q^*)^2}{2}$ (right panel).

□

Figure E.1 shows the results from numerical simulations compared with the theoretical prediction of Equation (E.14). The figure further illustrates that the model can generate two-vertex and three-vertex degree correlations, such as a decreasing average nearest neighbor connectivity, $k_{\text{nn}}(d)$, indicating a dissortative network, as well as a decreasing clustering degree distribution, $C(d)$, with the degree d.

E.2. Heterogeneous Technology Spillovers

In this section we allow for heterogeneity among firms in terms of their technological abilities [Griffith et al., 2003]. We assume that the technologies embodied in a firm $i \in \mathcal{N} = \{1, \ldots, n\}$ can be represented as an N-dimensional vector $\mathbf{h}_i$ in the technology space $\mathcal{H}^N = \{0, 1\}^N$, which consists of all binary sequences with elements in $\{0, 1\}$ of length N. The number of such sequences is 2^N. The technology vector $\mathbf{h}_i$, with components $h_{ik} \in \{0, 1\}$, indicates whether firm i knows idea $k \in \{1, \ldots, N\}$ or not. We introduce a spillover function $f : \mathcal{H}^N \times \mathcal{H}^N \to \mathbb{R}$ capturing the potential technology transfer between any pairs of firms. A possible specification is one in which $f(\mathbf{h}_i, \mathbf{h}_j) = \mathbb{1}_{\{\langle \mathbf{h}_i, \mathbf{h}_j \rangle > s\}}$, where $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in $\mathbb{R}^n$, so that $\langle \mathbf{h}_i, \mathbf{h}_j \rangle$ counts the number of technologies known to both i and j, and $f(\mathbf{h}_i, \mathbf{h}_j)$ is one iff i and j have at least $s > 0$ technologies in common. This is an instance of a *random intersection graph* [Deijfen and Kets, 2009] (see also Appendix B).[19]

[19] There is a variety of other functional forms that can be incorporated in our model. For example, a simple choice for the function f could be $f(\mathbf{h}_i, \mathbf{h}_j) = a|\mathbf{h}_i \cap \mathbf{h}_j|$, where $a \in \mathbb{R}_+$ and $|\mathbf{h}_i \cap \mathbf{h}_j| = \mathbf{h}_i^\top \mathbf{h}_j = \sum_{k=1}^N h_{ik} h_{jk}$ denotes the common knowledge of i and j. Alternative specifications for similarity can be found in Liben-Nowell and Kleinberg [2007] and Bloom et al. [2013]; Jaffe [1989]. Alternatively, following Berliant and Fujita [2008, 2009], a possible parametric specification for f would be $f(\mathbf{h}_i, \mathbf{h}_j) = |\mathbf{h}_i \cap \mathbf{h}_j|^\kappa d(\mathbf{h}_i, \mathbf{h}_j)^{\frac{1-\kappa}{2}}$ for some $\kappa \in (0, 1)$. The distance is the product of the total number of ideas known by agent i but not by j times the total number of ideas known by j but not by i, i.e. $d(\mathbf{h}_i, \mathbf{h}_j) = |\mathbf{h}_i \backslash \mathbf{h}_j| \times |\mathbf{h}_j \backslash \mathbf{h}_i| = |\mathbf{h}_i \cap \mathbf{h}_j^c| \times |\mathbf{h}_i^c \cap \mathbf{h}_j| = \sum_{k=1}^N h_{ik}(1 - h_{jk}) \sum_{k=1}^N (1 - h_{ik}) h_{jk}$, where $\mathbf{u} = (1, \ldots, 1)^\top$ and $\mathbf{h}_i^c = \mathbf{u} - \mathbf{h}_i$. Other functional forms have been suggested in the literature [see e.g., Baum et al., 2009; Nooteboom et al., 2007], such as $f(\mathbf{h}_i, \mathbf{h}_j) = a_1|\mathbf{h}_i \cap \mathbf{h}_j| - a_2|\mathbf{h}_i \cap \mathbf{h}_j|^2$, with constants $a_1, a_2 \geq 0$.

Given the spillover function $f(\mathbf{h}_i, \mathbf{h}_j)$, the marginal cost of production of a firm i becomes

$$c_i = \bar{c} - \alpha e_i - \beta \sum_{j=1}^{n} a_{ij} f(\mathbf{h}_i, \mathbf{h}_j) e_j,$$

and profits of firm i are given by

$$\pi_i = (a - \bar{c}) q_i - q_i^2 - b q_i \sum_{j \neq i} q_j + \alpha q_i e_i + \beta q_i \sum_{j=1}^{n} a_{ij} f(\mathbf{h}_i, \mathbf{h}_j) e_j - \gamma e_i^2 - \zeta d_i.$$

The optimal effort levels are given by $e_i = \frac{\alpha}{2\gamma} q_i = \lambda q_i$. Inserting into profits yields

$$\begin{aligned}
\pi_i &= (a - \bar{c}) q_i - (1 - \lambda\alpha + \lambda^2 \gamma) q_i^2 - b q_i \sum_{j \neq i} q_j + \lambda \beta q_i \sum_{j=1}^{n} a_{ij} f(\mathbf{h}_i, \mathbf{h}_j) q_j - \zeta d_i \\
&= \eta q_i - \nu q_i^2 - b q_i \sum_{j \neq i} q_j + \rho q_i \sum_{j=1}^{n} a_{ij} f(\mathbf{h}_i, \mathbf{h}_j) q_j - \zeta d_i.
\end{aligned}$$

We can then obtain a potential function (see Proposition 1) given by

$$\Phi(\mathbf{q}, G, \mathbf{h}) = \sum_{i=1}^{n} ((a - \bar{c}) q_i - \nu q_i^2) - \frac{b}{2} \sum_{i=1}^{n} q_i \sum_{j \neq i} q_j + \sum_{i=1}^{n} q_i \sum_{j=1}^{n} a_{ij} f(\mathbf{h}_i, \mathbf{h}_j) q_j - \zeta m.$$

The stationary distribution (see Theorem 1) is given by

$$\mu^{\vartheta}(\mathbf{q}, G, \mathbf{h}) = \frac{e^{\vartheta \Phi(\mathbf{q}, G, \mathbf{h})}}{\sum_{\mathbf{h}' \in \mathcal{H}^N} \sum_{G' \in \mathcal{G}^n} \int_{\mathcal{Q}^n} e^{\vartheta \Phi(\mathbf{s}, G', \mathbf{h}')} d\mathbf{s}}.$$

The probability of observing a network $G \in \mathcal{G}^n$, given an output distribution $\mathbf{q} \in [0, \overline{q}]^n$ and technology portfolios $\mathbf{h} \in \mathcal{H}^N$ is determined by the conditional distribution

$$\mu^{\vartheta}(G | \mathbf{q}, \mathbf{h}) = \prod_{i<j} \frac{e^{\vartheta a_{ij} (\rho f(\mathbf{h}_i, \mathbf{h}_j) q_i q_j - \zeta)}}{1 + e^{\vartheta (\rho f(\mathbf{h}_i, \mathbf{h}_j) q_i q_j - \zeta)}}, \tag{E.17}$$

which is equivalent to the probability of observing an inhomogeneous random graph with link probability

$$p^{\vartheta}(q_i, \mathbf{h}_i, q_j, \mathbf{h}_j) \equiv \frac{e^{\vartheta (\rho f(\mathbf{h}_i, \mathbf{h}_j) q_i q_j - \zeta)}}{1 + e^{\vartheta (\rho f(\mathbf{h}_i, \mathbf{h}_j) q_i q_j - \zeta)}}. \tag{E.18}$$

In the following we consider a particularly simple specification in which each firm i is assigned a technology $k \in \{1, \ldots, N\}$ uniformly at random so that $h_{ik} = 1$ and $h_{il} = 0$ for all $l \neq k$. Moreover, let $f(\mathbf{h}_i, \mathbf{h}_j) = \mathbb{1}_{\{\langle \mathbf{h}_i, \mathbf{h}_j \rangle \geq 1\}}$, that is, firms i an j can only benefit from a collaboration if they have a technology in common.

Proposition 8. *Assume that each firm i is assigned a technology $k \in \{1, \ldots, N\}$ uniformly at random and let $f(\mathbf{h}_i, \mathbf{h}_j) = \mathbb{1}_{\{\langle \mathbf{h}_i, \mathbf{h}_j \rangle \geq 1\}}$.*

(i) The degree distribution is given by

$$P(k) = \binom{n}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{n-k}.$$

(ii) The average nearest neighbor degree distribution is given by

$$k_{nn}(k) = \frac{k(1-\frac{1}{N})(1+n\frac{1}{N}-(n+1)\left(\frac{1}{N}\right)^n)}{\frac{1}{N}(1+n-k)}, \tag{E.19}$$

and for large n the average nearest neighbor degree distribution, $k_{nn}(k)$, grows linearly as $O(k)$.

(iii) The clustering coefficient is given by $C(k) = 1$.

Proof of Proposition 8. We first prove part (i) of the proposition. If technologies are assigned uniformly at random then

$$\mathbb{P}\left(\langle \mathbf{h}_i, \mathbf{h}_j\rangle \geq 1 \middle| q_i = q, q_j = q'\right) = \frac{1}{N}\mathbb{1}_{\{\rho q q' > \zeta\}}.$$

Due to symmetry the firms quantities in the stationary state when $\vartheta \to \infty$ are identical and given by q^*. In the case of $\rho(q^*)^2 > \zeta > 0$ we then we have that

$$\mathbb{P}\left(a_{ij} = 1\right) = \frac{1}{N},$$

and the degree distribution is given by

$$P(k) = \mathbb{P}\left(d_1(G) = k\right) = \binom{n}{k}\left(\frac{1}{N}\right)^k\left(1-\frac{1}{N}\right)^{n-k}.$$

We next give a proof of part (ii) of the proposition. The average nearest neighbor degree distribution is then given by

$$k_{\rm nn}(k) = \sum_{k'=1}^{n} k'\mathbb{P}\left(d_2(G) = k'-1 \middle| a_{12} = 1, d_1(G) = k\right)$$

where

$$\begin{aligned}\mathbb{P}\left(d_2(G) = k'-1 \middle| a_{12} = 1, d_1(G) = k\right) &= \frac{\mathbb{P}\left(d_2(G) = k'-1, d_1(G) = k \middle| a_{12} = 1\right)}{P(k)} \\ &= \frac{1}{P(k)}\left(\frac{1}{N}\right)^{k'-1}\left(1-\frac{1}{N}\right)^{n-k'+1}\left(\frac{1}{N}\right)^{k-1}\left(1-\frac{1}{N}\right)^{n-k+1} \\ &= \frac{P(k'-1)P(k-1)}{P(k)}.\end{aligned}$$

We then get

$$\begin{aligned}k_{\rm nn}(k) &= \sum_{k'=1}^{n} k'\mathbb{P}\left(d_2(G) = k'-1 \middle| a_{12} = 1, d_1(G) = k\right) \\ &= \sum_{k'=1}^{n} k'\frac{P(k'-1)P(k-1)}{P(k)} \\ &= \frac{k(1-\frac{1}{N})(1+n\frac{1}{N}-(n+1)\left(\frac{1}{N}\right)^n)}{\frac{1}{N}(1+n-k)} \\ &= O(k),\end{aligned}$$

as $n \to \infty$. That is, the average nearest neighbor degree $k_{\rm nn}(k)$ is asymptotically linearly increasing with the degree k, and thus we have an assortative network.

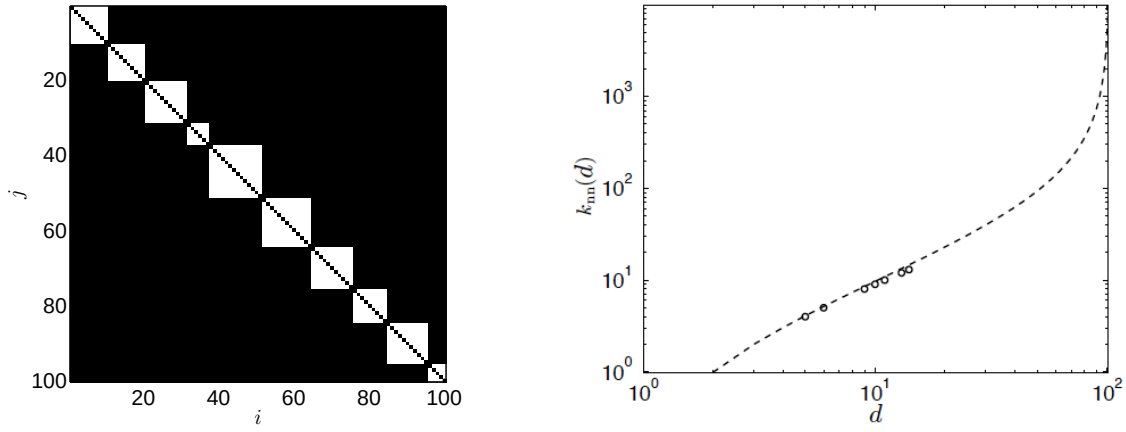


Figure E.3: (Left panel) Illustration of the matrix with elements $\mathbb{1}_{\{\langle \mathbf{h}_i, \mathbf{h}_j \rangle > 0\}}$ for $n = 100$ firms and $N = 10$ technologies. (Right panel) The average nearest neighbor degree distribution, $k_{\rm nn}(d)$, for the same parameters. The dashed line represents the solution from Equation (E.19) while the circles correspond to a numerical simulation.

Finally, we give a proof of part (iii) of the proposition. The clustering coefficient is simply given by $C(k) = \mathbb{P}\left(a_{23} = 1 | \, a_{12} = 1, a_{23} = 1, d_1(G) = k\right) = 1$. This is because if firm 1 is connected to firm 2 then they must have the same technology. Similarly, if firm 1 is connected to firm 3 then they also must have the same technology. Due to transitivity, firms 2 and 3 then must have the same technology, and thus must be connected. □

An illustration of the average nearest neighbor degree $k_{\rm nn}(k)$ can be seen in Figure E.3.

F. Data

In the following we provide a detailed description of the data used for our empirical analysis in Section 3.

To get a comprehensive picture of alliances we use data on interfirm R&D collaborations stemming from two sources which have been widely used in the literature [Schilling, 2009]. The first is the Cooperative Agreements and Technology Indicators (CATI) database [Hagedoorn, 2002]. The database only records agreements for which a combined innovative activity or an exchange of technology is at least part of the agreement. Moreover, only agreements that have at least two industrial partners are included in the database, thus agreements involving only universities or government labs, or one company with a university or lab, are disregarded. The second is the Thomson Securities Data Company (SDC) alliance database. SDC collects data from the U. S. Securities and Exchange Commission (SEC) filings (and their international counterparts), trade publications, wires, and news sources. We include only alliances from SDC which are classified explicitly as research and development collaborations. A comparative analysis of these two databases (and other alternative databases) can be found in Schilling [2009].

We merged the CATI database with the Thomson SDC alliance database. For the matching of firms across datasets, we adopted the name matching algorithm developed as part of the NBER patent data project [Trajtenberg et al., 2009]. We could match 21% of the firms appearing in both databases. Considering only firms without missing observations on R&D expenditures and industry classifications (see also Appendix F.2 below on how we obtained balance sheet and R&D expenditures information), it gives us a sample of $2,033$ firms and a total of 720 collaborations in the year 2006. The average degree of the firms in this sample is 0.71 with a standard deviation of 2.01, and the maximum degree is 25 attained by *Pfizer Inc.*

Figure F.1, and Tables F.1 and F.2 show the 10, respectively 20, largest sectors at the 2-digit and 3-digit SIC levels. The largest sector at the SIC-28 level is chemical and allied products, with 534 firms (26.27 % of the total), followed by the sector electronic and other electric equipment, with 306 firms (15.05 % of the total). At the 3-digit SIC level the largest sector is the drugs development sector, with 416 firms (20.46 % of the total), and the second largest sector is computer and data processing services with 193 firms (9.49 % of the total).

Figure F.2 shows the degree distribution, $P(d)$, the average nearest neighbor connectivity, $k_{\text{nn}}(d)$, the clustering degree distribution, $C(d)$, and the component size distribution, $P(s)$ across different levels of sectoral aggregation, considering all firms in all sectors, firms in the SIC-28 sector only, or firms in the SIC-283 sector only. The degree distribution, $P(d)$, decays as a power law across all datasets considered. The clustering degree distribution, $C(d)$, is also decreasing with increasing degrees d across all datasets. These network tend to be moderately clustered. The average clustering coefficient considering all firms is $C = 0.074$, for the firms in the SIC-28 sector it is $c = 0.043$ and for the firms in the SIC-283 sector it is $C = 0.038967$. Further, the component size distribution, $P(s)$, indicates a large connected component (see also Figure 7) with smaller components decaying as a power law. This pattern is also consistent across datasets. The largest connected component comprises 21.20% of all firms across sectors, $24,07$% of all firms in the SIC-28 sector, and 29.91% of all firms in the SIC-283 sector. While the level or sectoral aggregation does not matter much for the degree distribution, the clustering degree distribution and the component size distribution, a different pattern can be observed for the average nearest neighbor connectivity, $k_{\text{nn}}(d)$. While the average nearest neighbor connectivity

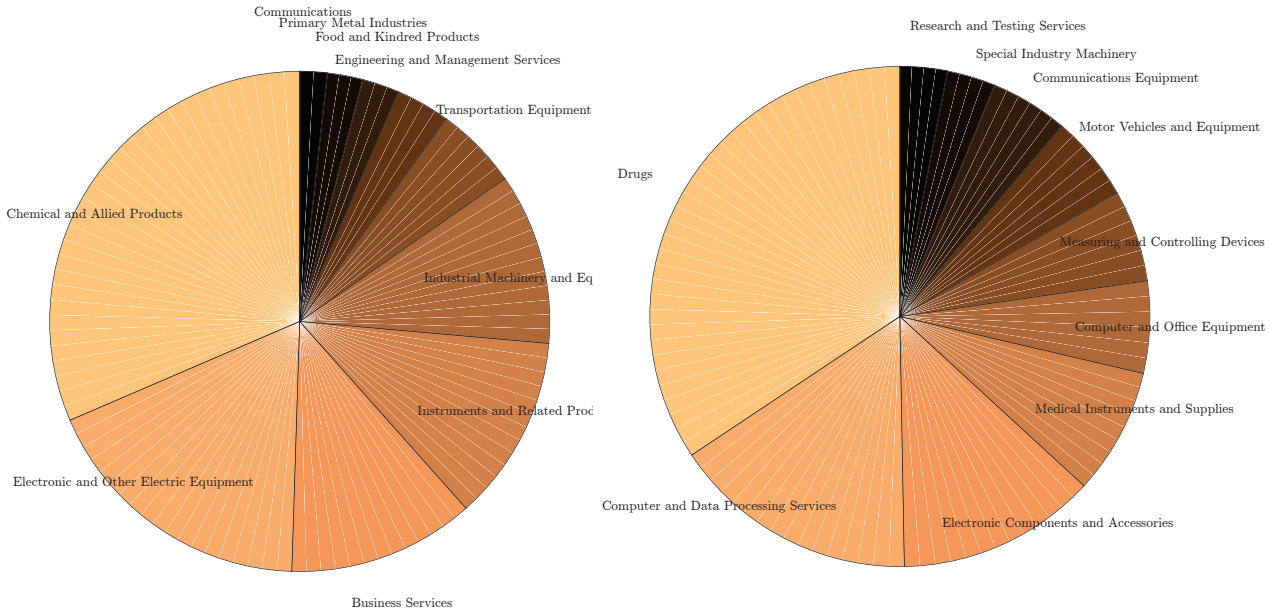


Figure F.1: The shares of the ten largest sectors at the 2-digit (left panel) and 3-digit (right panel) SIC levels. See also Tables F.1 and F.2, respectively.

Table F.1: The 20 largest sectors at the 2-digit SIC level.

Sector	2-dig SIC	# firms	% of tot.	Rank
Chemical and Allied Products	28	534	26.27	1
Electronic and Other Electric Equipment	36	306	15.05	2
Business Services	73	205	10.08	3
Instruments and Related Products	38	204	10.03	4
Industrial Machinery and Equipment	35	188	9.25	5
Transportation Equipment	37	91	4.48	6
Engineering and Management Services	87	59	2.90	7
Food and Kindred Products	20	41	2.02	8
Primary Metal Industries	33	39	1.92	9
Communications	48	30	1.48	10
Electric Gas and Sanitary Services	49	24	1.18	11
Fabricated Metal Products	34	21	1.03	12
Health Services	80	20	0.98	13
Miscellaneous Manufacturing Industries	39	18	0.89	14
Paper and Allied Products	26	16	0.79	15
Rubber and Miscellaneous Plastics Products	30	16	0.74	16
Petroleum and Coal Products	29	14	0.69	17
Stone Clay and Glass Products	32	14	0.69	18
Wholesale Trade - Durable Goods	50	13	0.64	19
Textile Mill Products	22	12	0.59	20

Table F.2: The 20 largest sectors at the 3-digit SIC level.

Sector	3-dig SIC	# firms	% of tot.	Rank
Drugs	283	416	20.46	1
Computer and Data Processing Services	737	193	9.49	2
Electronic Components and Accessories	367	156	7.67	3
Medical Instruments and Supplies	384	99	4.87	4
Computer and Office Equipment	357	72	3.54	5
Measuring and Controlling Devices	382	72	3.54	6
Motor Vehicles and Equipment	371	67	3.30	7
Communications Equipment	366	61	3.00	8
Special Industry Machinery	355	38	1.87	9
Research and Testing Services	873	37	1.82	10
Misc. Electrical Equipment and Supplies	369	28	1.38	11
Chemicals & Allied Products	280	26	1.28	12
Plastics Materials and Synthetic	282	25	1.23	13
General Industrial Machinery	356	25	1.23	14
Electrical Industrial Apparatus	362	17	0.84	15
Aircraft and Parts	372	16	0.79	16
Misc. Chemical Products	289	15	0.74	17
Blast Furnace and Basic Steel Products	331	15	0.74	18
Agricultural Chemicals	287	14	0.69	19
Metalworking Machinery	354	14	0.69	20

$k_{\text{nn}}(d)$ is decreasing with increasing degree for the firms restricted to the SIC-28 or SIC-283 sectors, this monotonicity behavior is less pronounced when considering all firms across sectors. This pattern can also be observed in the assortativity coefficient, which is $\gamma = -0.031399$ for all firms, $\gamma = -0.25322$ restricting the sample to firms in the SIC-28 sector, and $\gamma = -0.27464$ for the firms in the SIC-283 sector. That is, while the network is weakly dissortative considering all firms, it becomes strongly dissortative when considering only a single sector. This observation is even more extreme when we consider all firms in the collaboration network without dropping those for which R&D expenditures are missing. In this case we find $\gamma = 0.03343$ for all firms, $\gamma = -0.11703$ restricting the sample to firms in the SIC-28 sector, and $\gamma = -0.14886$ for the firms in the SIC-283 sector. The explanation for this observation can be easily given when considering the extension of our model introduced in Section 2.4 and Appendix E.2. There we showed that when the spillovers from collaborations depend on the technological characteristics of the firms involved in a collaboration, and firms from different sectors have different characteristics, then the emerging network of cross industry collaborations can be assortative, while the network of intra-industry collaborations is dissortative.

F.1. Mergers and Acquisitions

Some firms might be acquired by other firms due to mergers and acquisitions (M&A) over time, and this will impact the R&D collaboration network [Hanaki et al., 2010].

To get a comprehensive picture of the M&A activities of the firms in our dataset, we use two extensive datasources to obtain information about M&As. The first is the Thomson Reuters' Securities Data Company (SDC) M&A database, which has historically been the most widely used database for empirical research in the field of M&As. Data in SDC dates back to 1965 with a slightly more complete coverage of deals starting in the early 1980s. The second database with information about M&As is Bureau van Dijk's (BvD) Zephyr database, which is a recent alternative to the SDC M&As database. The history of deals recorded in Zephyr goes back to

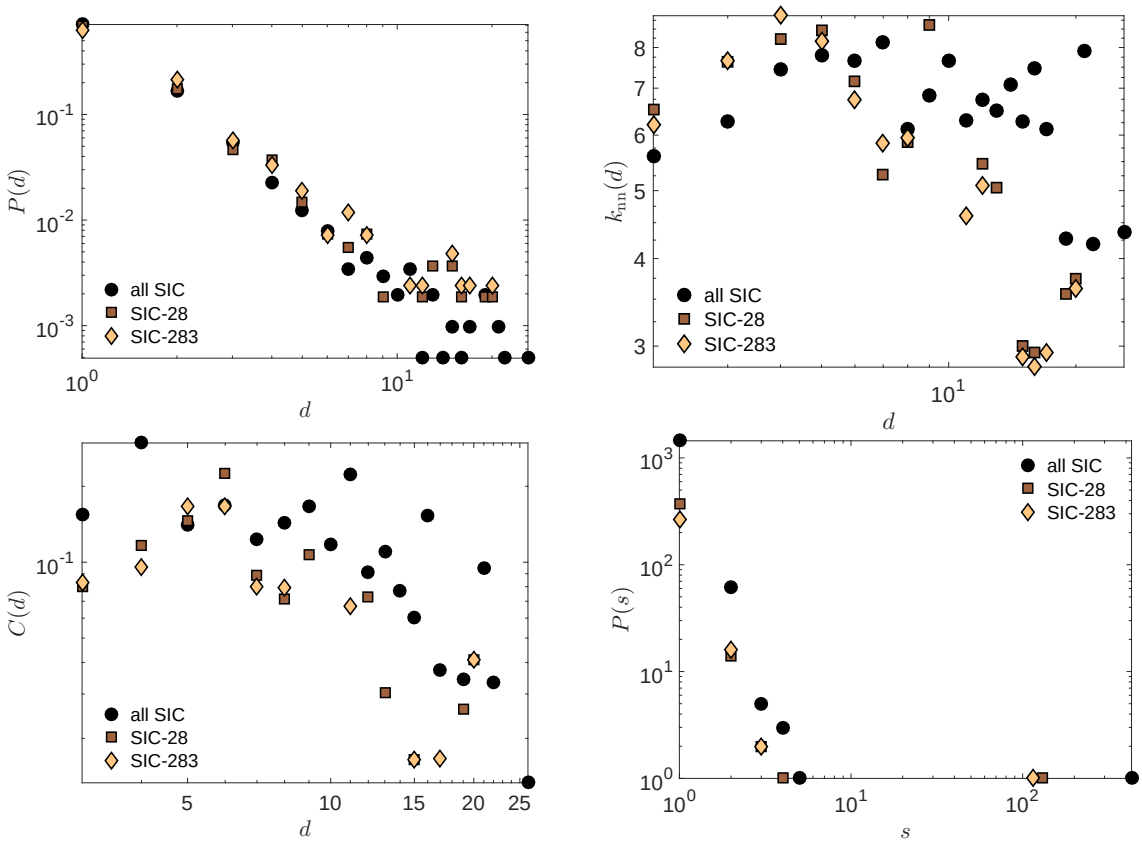


Figure F.2: The degree distribution, $P(d)$, the average nearest neighbor connectivity, $k_{\rm nn}(d)$, the clustering degree distribution, $C(d)$, and the component size distribution, $P(s)$.

1997. In 1997 and 1998 only European deals are recorded, while international deals are included starting from 1999. According to Huyghebaert and Luypaert [2010], Zephyr "covers deals of smaller value and has a better coverage of European transactions". A comparison and more detailed discussion of the two databases can be found in Bollaert and Delanghe [2015] and Bena et al. [2008].

We merged the SDC and Zephyr databases (with the above mentioned name matching algorithm; see also Trajtenberg et al. [2009]) to obtain information on M&As of 116,641 unique firms. Using the same name matching algorithm we could identify 43.08% of the firms in the combined CATI-SDC alliance database that also appear in the combined SDC-Zephyr M&As database. We then account for the M&A activities of these matched firms when constructing the R&D collaboration network by assuming that an acquiring firm in a M&A inherits all the R&D collaborations of the target firm, and we remove the target firm form from the network.

F.2. Balance Sheet Statements, R&D and Productivity

The combined CATI-SDC alliance database provides the names for each firm in an alliance, but it does not contain information about the firms' output levels or R&D expenses. We therefore matched the firms' names in the combined CATI-SDC database with the firms' names in Standard & Poor's Compustat U.S. and Global fundamentals annual databases and Bureau van Dijk (BvD)'s Orbis database, to obtain information about their balance sheets and income statements.

Compustat North America is a database of U.S. and Canadian active and inactive publicly held companies extracted from company filings. It provides more than 300 annual and 100 quarterly income statements, balance sheets and statement of cash flows. Compustat Global is

a database of non-U.S. and non-Canadian companies and contains market information on more than 33,900 active and inactive publicly held companies with annual data history from 1987. The Compustat databases cover 99% of the world's total market capitalization with annual company data history available back to 1950. The databases contains only firms listed on the stock market, so it typically excludes smaller private firms, but this is inevitable if one is going to use market value data. Nevertheless, R&D is concentrated in publicly listed firms, and it thus covers most of the R&D activities in the economy [Bloom et al., 2013].

The Orbis database is owned by Bureau van Dijk (BvD). It is a commercial dataset, which contains administrative data on 130 million firms worldwide. Orbis is an umbrella product that provides firm-level data covering over 120 countries, both developed and emerging, since 2005. The financial and balance-sheet information in Orbis comes from business registers collected by the local Chambers of Commerce to fulfil legal and administrative requirements and are relayed to BvD via over 40 different information providers. Different from Compustat, Orbis contains not only information about publicly listed firms, but provides also information about private firms.

For a detailed comparison and further discussion of the Compustat and Orbis databases see Dai [2012], Bloom et al. [2013] and Papadopoulos [2012].

For the matching of firms across datasets we adopted the name matching algorithm developed as part of the NBER patent data project [Trajtenberg et al., 2009]. We could match 25.53% of the firms in the combined CATI-SDC database with the combined Compustat-Orbis database. For the matched firms we obtained their their sales, R&D expenditures, sales, employment, primary industry codes and location. U.S. dollar translation rates for foreign currencies have been taken directly from the Compustat yearly averaged exchange rates. We adjusted for inflation using the consumer price index of the Bureau of Labor Statistics (BLS), averaged annually, with 1983 as the base year. From a firm's sales and employment we then computed its labor productivity as sales relative to the number of employees. We then dropped all firms with missing information on R&D expenditures and industry codes. This pruning procedure left us with a subsample of $2,033$ firms, on which the empirical analysis in Section 3 is based.

The empirical distributions for sales, $P(s)$, productivity, $P(x)$, R&D expenditures, $P(e)$, and the patent stocks, $P(k)$ (using a logarithmic binning of the data with 100 bins [McManus et al., 1987]) are shown in Figure F.3. All distributions are highly skewed, indicating a large degree of inequality in firms' sizes, productivity and patent activities. Moreover, Figure F.4 shows a correlation scatter plot for sales, productivity, R&D expenditures and the patent stocks. All are highly correlated, with a Spearman correlation coefficient between sales and R&D expenditures of $\varrho = 0.69$ between sales and productivity of $\varrho = 0.54$, and sales and the number of patents of $\varrho = 0.53$. The correlation between R&D expenditure and productivity is $\varrho = 0.29$ and R&D expenditures and the number of patents is $\varrho = 0.56$. Finally, the correlation between productivity and the number of patents is $\varrho = 0.22$.

F.3. Geographic Location and Distance

The number of firms in each country is shown in Figures F.5 and F.6, respectively, while Table F.3 shows the 25 countries with the largest numbers of firms. The dominant role of the U.S. with 989 collaborations making up 48.65% of the total number of collaborations is clearly visible. The second largest country in terms of R&D collaborations is Japan with 408, which comprises 20.07% of all collaborations. The U.S. and Japan then together account for 68.72%, that is, more

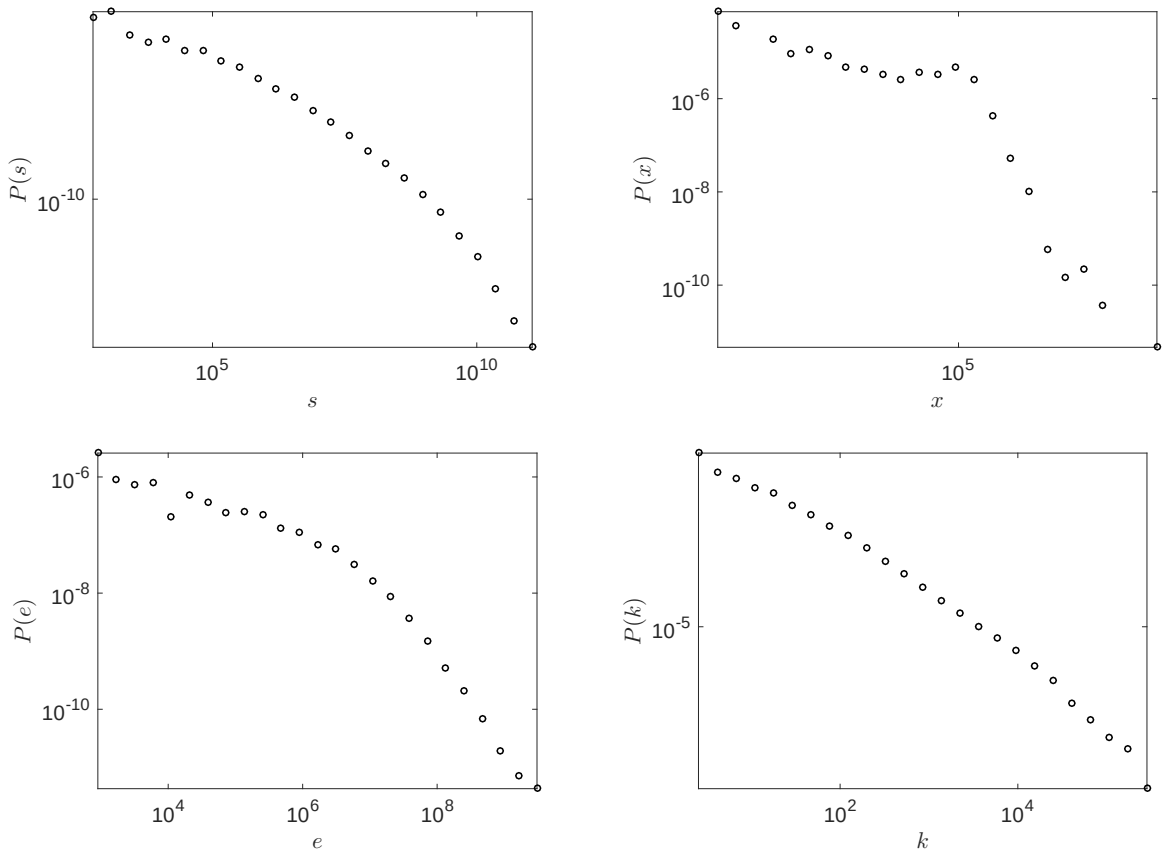


Figure F.3: The sales distribution, $P(s)$, the productivity distribution, $P(x)$, the R&D expenditures distribution, $P(e)$, and the patent stock distribution, $P(k)$ using a logarithmic binning of the data [McManus et al., 1987].

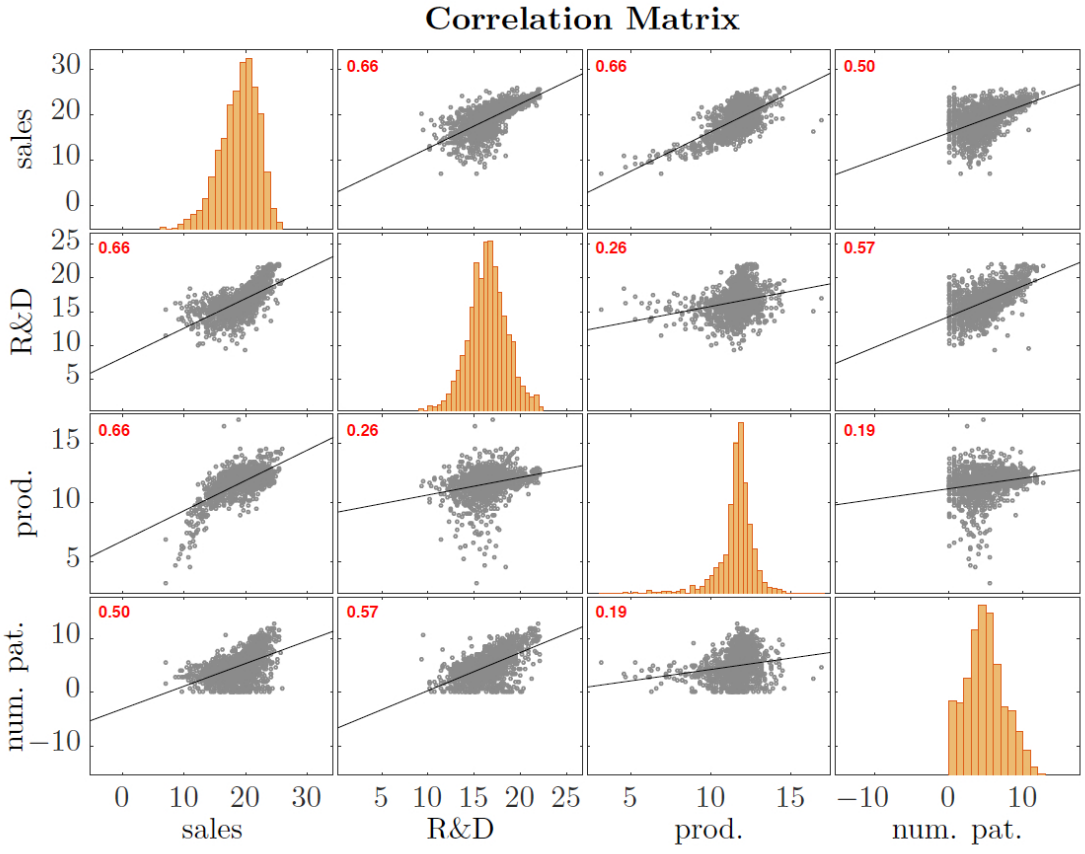


Figure F.4: Correlation scatter plot for sales, productivity, R&D expenditures and the patent stocks.

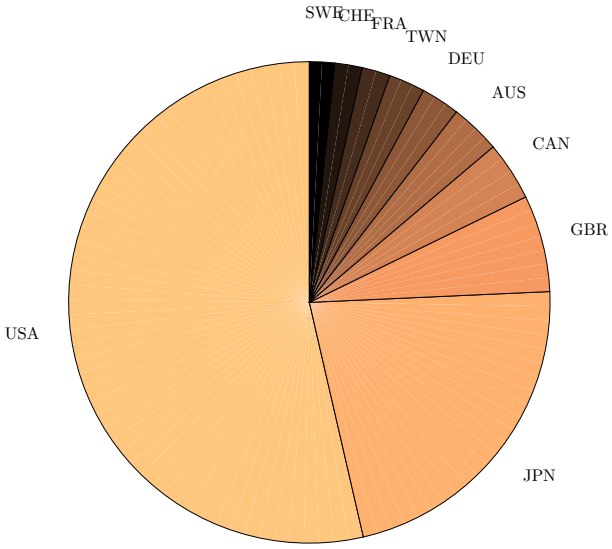


Figure F.5: The number of firms in each country.

than two thirds of all R&D collaborations in the data.

In order to determine the precise locations of the firms in our data we have further added the longitude and latitude coordinates associated with the city of residence of each firm. Among the matched cities in our dataset 93.67% could be geo-localized using ArcGIS [see e.g., Dell, 2009] and the Google Maps Geocoding API.[20] We then used Vincenty's algorithm to compute the distances between pairs of geo-localized firms [Vincenty, 1975]. The mean distance between collaborating firms is $5,227$ km. The distance distribution, $P(d)$, across collaborating firms is shown in Figure F.8, while Figure F.7 shows the locations (at the city level) and collaborations of the firms in the database. The distance distribution, $P(d)$, is heavily skewed. We find that R&D collaborations tend to be more likely between firms that are close, showing that geography matters for R&D collaborations and this spillovers, in line with previous empirical studies [Lychagin et al., 2016].

F.4. Patents and Technological Similarity

We identified the patent portfolios of the firms in our dataset using the EPO Worldwide Patent Statistical Database (PATSTAT) [Jaffe and Trajtenberg, 2002; Thoma et al., 2010]. It includes bibliographic details on patents filed to 80 patent offices worldwide, covering more than 60 million documents. Hence filings in all major countries and the the World International Patent Office are covered. We matched the firms in our data with the assignees in the PATSTAT database using the above mentioned name matching algorithm. We only consider granted patents (or successful patents), as opposed to patents applied for, as they are the main drivers of revenue derived from R&D [Copeland and Fixler, 2012]. We obtained matches for roughly 30% of the firms in the data. The technology classes were identified using the main international patent classification (IPC) numbers at the 4 digit level.

[20] See https://developers.google.com/maps/documentation/geocoding/intro.

Table F.3: The 25 countries with the largest numbers of firms.

Name	Code	# firms	% of tot.	Rank
United States	USA	989	48.65	1
Japan	JPN	408	20.07	2
United Kingdom	GBR	120	5.90	3
Canada	CAN	73	3.59	4
Australia	AUS	62	3.05	5
Germany	DEU	48	2.36	6
Taiwan	TWN	45	2.21	7
France	FRA	35	1.72	8
Switzerland	CHE	34	1.67	9
Sweden	SWE	31	1.52	10
India	IND	21	1.03	11
Finland	FIN	16	0.79	12
Netherlands	NLD	14	0.69	13
Iceland	ISL	13	0.64	14
Slovakia	SVK	13	0.64	15
Denmark	DNK	12	0.59	16
Belgium	BEL	11	0.54	17
Italy	ITA	11	0.54	18
Israel	ISR	10	0.49	19
Morocco	MAR	9	0.44	20
Norway	NOR	8	0.39	21
China	CHN	7	0.34	22
Singapore	SGP	7	0.34	23
Spain	ESP	4	0.20	24
Hong Kong	HKG	4	0.20	25

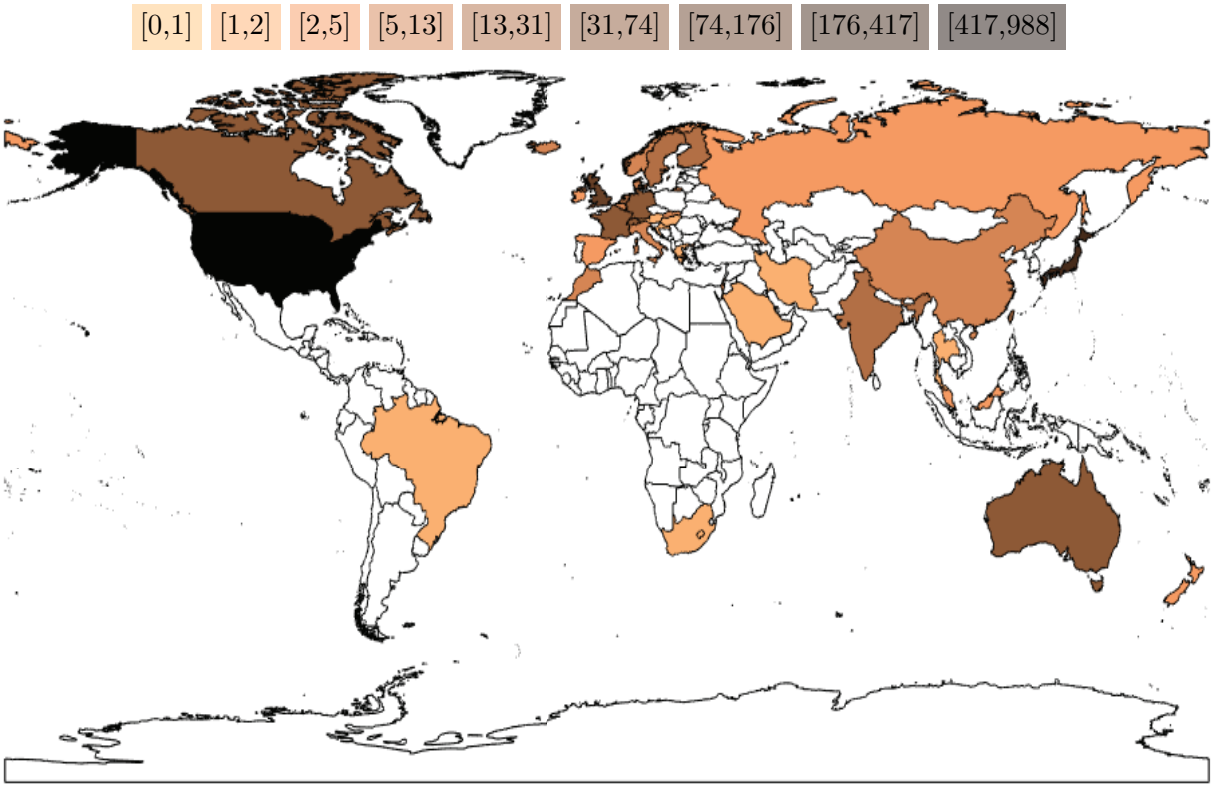


Figure F.6: The number of firms in each country.

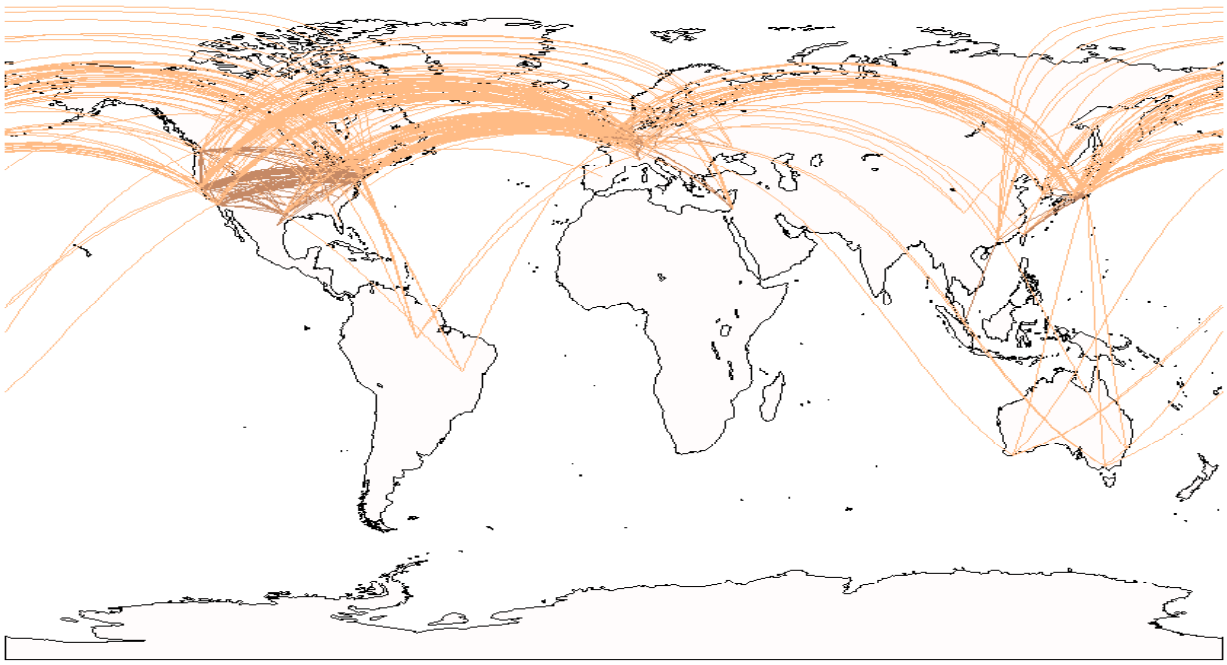


Figure F.7: The locations (at the city level) and collaborations of the firms in the combined CATI-SDC database.

The technological proximity $f_{ij}^k = f_{ji}^k$, $k \in \{J, M\}$ (see Equation (23)), between firms i and j is measured with two alternative metrics. The first, f_{ij}^J, is based on Jaffe [1989]. Let $\mathbf{P}_i$ represents the patent portfolio of firm i, where, for each firm i, $\mathbf{P}_i$ is a vector whose k-th component, P_{ik}, counts the number of patents firm i has in technology category k divided by the total number of technologies attributed to the firm [see also Bloom et al., 2013]. The technological proximity of firm i and j is then given by

$$f_{ij}^J = \frac{\mathbf{P}_i^\top \mathbf{P}_j}{\sqrt{\mathbf{P}_i^\top \mathbf{P}_i}\sqrt{\mathbf{P}_j^\top \mathbf{P}_j}}. \tag{F.20}$$

We denote $\mathbf{F}^J$ the $(n \times n)$ matrix with elements $(f_{ij}^J)_{1 \le i,j \le n}$.

As an alternative measure for technological similarity we also consider the Mahalanobis technology proximity measure, f_{ij}^M, introduced by Bloom et al. [2013]. To construct this metric, let N be the number of technology classes, n the number of firms, and let $\mathbf{T}$ be the $(N \times n)$ patent shares matrix with elements $T_{ji} = P_{ji} / \sum_{k=1}^n P_{ki}$, for all $1 \le i \le n$ and $1 \le j \le N$. Further, we construct the $(N \times n)$ normalized patent shares matrix $\tilde{\mathbf{T}}$ with elements $\tilde{T}_{ji} = T_{ji} / \sqrt{\sum_{k=1}^N T_{ki}^2}$, and the $(n \times N)$ normalized patent shares matrix across firms is defined by $\tilde{\mathbf{X}}$ with elements $\tilde{X}_{ik} = T_{ki} / \sqrt{\sum_{i=1}^N T_{ki}^2}$. Let $\mathbf{\Omega} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$. Then the $(n \times n)$ Mahalanobis technology similarity matrix $\mathbf{F}^M = (f_{ij}^M)_{1 \le i,j \le n}$ is defined as

$$\mathbf{F}^M = \tilde{\mathbf{T}}^\top \mathbf{\Omega} \tilde{\mathbf{T}}. \tag{F.21}$$

We then use either f_{ij}^J or f_{ij}^M as a measure for the potential technology spillovers between collaborating firms in the profit function of Equation (23). Both measures are highly correlated. The Spearman correlation coefficient between the Jaffe and the Mahalanobis proximity metrics is 0.91882, and a correlation plot can be seen in the right panel of Figure F.8.

G. Estimation Algorithms

In the following appendices we provide additional details regarding our estimation algorithms.

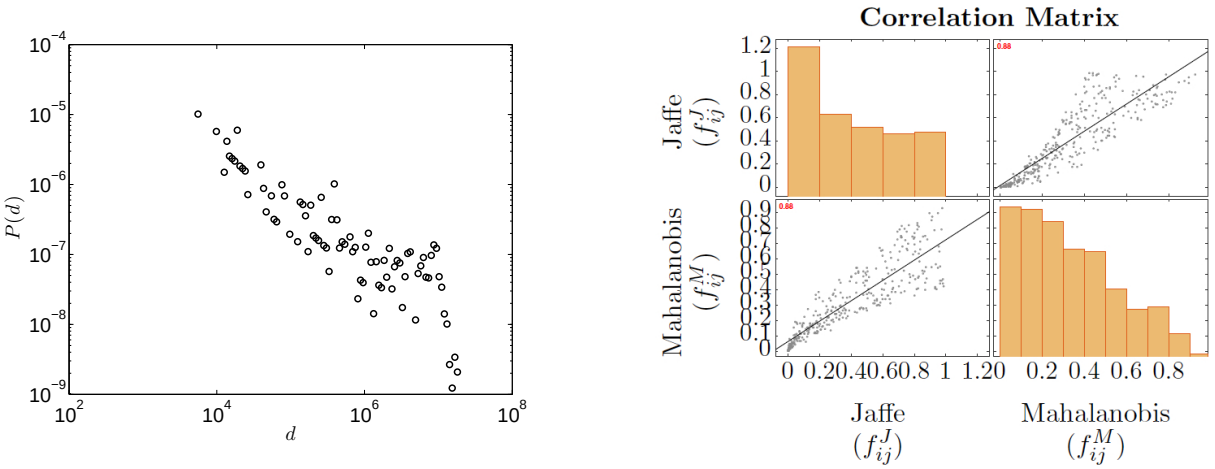


Figure F.8: (Left panel) The distance distribution, $P(d)$, across collaborating firms in the combined CATI-SDC database. (Right panel) Correlation plot for the Jaffe (f_{ij}^J) and the Mahalanobis (f_{ij}^M) technology proximity metrics across pairs of firms $1 \leq i, j \leq n$.

First, in Appendix G.1 we give a detailed explanation for how the DMH algorithm is implemented. Second, in Appendix G.2 we discuss the implementation of the AEX algorithm. Further, in Appendix G.3 we outline the convergence proof of the AEX algorithm.

G.1. Implementation of the DMH Algorithm

In our empirical model, the unknown parameters are denoted by $\boldsymbol{\theta} = (\rho, b, \boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top, \varkappa)$, where ρ captures the technology spillover effect; b captures the competition effect; $\boldsymbol{\delta}$ captures effects of firm's exogenous characteristics in terms of productivity and sector fixed effect; $\boldsymbol{\gamma}$ captures effects of dyad-specific exogenous factors in the link cost function; and $\varkappa$ captures the cyclic triangles effect. For the Bayesian posterior analysis, the joint posterior (probability) density function of $\boldsymbol{\theta}$ can be constructed by

$$P(\boldsymbol{\theta}|\mathbf{q}, G) \propto \pi(\boldsymbol{\theta}) \cdot \mu(\mathbf{q}|G, \boldsymbol{\theta}), \tag{G.22}$$

where $\pi(\cdot)$ represents the prior density function and we assume independence between prior distributions. We specify prior distributions $(\rho, b) \sim \mathscr{U}_2(\mathcal{O})$, $\boldsymbol{\delta} \sim \mathscr{N}_l(\boldsymbol{\delta}_0, \Delta_0)$, $\boldsymbol{\gamma} \sim \mathscr{N}_r(\boldsymbol{\gamma}_0, \Gamma_0)$, and $\varkappa \sim \mathscr{N}(\varkappa_0, \mathbf{K}_0)$. The prior of (ρ, b) is a multivariate uniform (uninformative) distribution with the values restricted to the compact subset $\mathcal{O}$ of $\mathbb{R}^2$, in which the matrix $\mathbf{M}(G) = \mathbf{I}_n + b\mathbf{B} - \rho(\mathbf{A} \circ \mathbf{F})$ is positive definite.[21] The priors for $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$, and $\varkappa$ are normal distributions, which are conjugate priors commonly used in the Bayesian literature [Koop et al., 2007; Robert and Casella, 2004]. In our applications, we fix $\boldsymbol{\delta}_0 = 0$, $\Delta_0 = 100\mathbf{I}_l$, $\boldsymbol{\gamma}_0 = 0$, $\Gamma_0 = 100\mathbf{I}_r$, $\varkappa_0 = 0$, and $\mathbf{K}_0 = 100$ to ensure prior density is relatively flat over the range of the data.

Directly drawing samples from the posterior distribution, $P(\boldsymbol{\theta}|\mathbf{q}, G)$, would be difficult due to high dimensionality. Therefore, we apply the Metropolis-within-Gibbs sampling procedure: we first divide unknown parameters into blocks so that drawing from their conditional posterior

[21] This requirement implies that $\mathbf{M}(G)$ is invertible and $\mathbf{M}(G)^{-1}$ is also positive definite. $\mathbf{M}(G)$ is positive definite if and only if all of its eigenvalues are positive.

distributions is feasible. When we make draws sequentially from these conditional posterior distributions by cycling, in the limit these draws can be treated as draws from the joint posterior distribution. For the conditional posterior distribution of (ρ, b), which is not available in a closed form, we use the MH algorithm to draw from this conditional distribution. It has been shown in Tierney [1994] and Chib and Greenberg [1996] that the combination of Markov chains is still a Markov chain with the invariant distribution equal to the correct objective distribution.

We apply the Gibbs sampler to update the following set of conditional posterior distributions:

(i) $P(\rho, b|\mathbf{q}, G, \boldsymbol{\theta}\backslash(\rho, b)) \propto \mu(\mathbf{q}, G|\boldsymbol{\theta}) I((\rho, b) \in O)$, where $\boldsymbol{\theta}\backslash(\rho, b)$ stands for $\boldsymbol{\theta}$ with (ρ, b) excluded.

(ii) $P(\boldsymbol{\delta}|\mathbf{q}, G, \boldsymbol{\theta}\backslash\boldsymbol{\delta}) \propto \phi_l(\boldsymbol{\delta}; \boldsymbol{\delta}_0, \Delta_0) \cdot \mu(\mathbf{q}|G, \boldsymbol{\theta})$, where ϕ denotes the normal density function.

(iii) $P(\boldsymbol{\gamma}|\mathbf{q}, G, \boldsymbol{\theta}\backslash\boldsymbol{\gamma}) \propto \phi_r(\boldsymbol{\gamma}; \boldsymbol{\gamma}_0, \Gamma_0) \cdot \mu(\mathbf{q}, G|\boldsymbol{\theta})$.

(iv) $P(\varkappa|\mathbf{q}, G, \boldsymbol{\theta}\backslash\varkappa) \propto \phi(\varkappa; \varkappa_0, \mathbf{K}_0) \cdot \mu(\mathbf{q}, G|\boldsymbol{\theta})$.

We apply the DMH algorithm to simulate draws from (i) to (iv) because their distributions are not available in a closed form. At each t^{th} MCMC iteration, the implementation steps are illustrated as follows:

Step I. Simulate $(\rho^{(t)}, b^{(t)})$ from $P\left(\rho, b \left| \mathbf{q}, G, \Upsilon^{(t-1)} \right.\right)$ by the DMH algorithm, where $\Upsilon^{(t-1)}$ denotes the rest of parameters evaluated at the $(t-1)^{\text{th}}$ iteration.

(a) propose $(\widetilde{\rho}, \widetilde{b})$ from a random walk proposal density $T_1(\rho, b|b^{(t-1)}, \rho^{(t-1)})$.

(b) simulate an auxiliary data $\mathbf{q}'$ and G' by M runs of the MH algorithm based on

$$\mu(\mathbf{q}, G|\widetilde{\rho}, \widetilde{b}, \Upsilon^{(t-1)}) = \frac{\exp\left(\Phi\left(\mathbf{q}, G, \widetilde{\rho}, \widetilde{b}, \Upsilon^{(t-1)}\right)\right)}{\sum_{\mathfrak{g}\in\mathcal{G}^n} \int_{\mathcal{Q}^n} \exp\left(\Phi\left(\mathfrak{q}, \mathfrak{g}, \widetilde{\rho}, \widetilde{b}, \Upsilon^{(t-1)}\right)\right) d\mathfrak{q}}$$

starting from the observed network G.[22] The details are as follows:
First, let the initial auxiliary network $G^{(0)}$ equal to the observed network G, i.e., set the auxiliary adjacency matrix $\mathbf{A}^{(0)} = \mathbf{A}$. At the m^{th} run, $m = 1, \cdots, M$, we propose $\widetilde{G}$ from $G^{(m-1)}$ by selecting node i with probability $\frac{1}{n}$ and node $j \neq i$ with a probability $\frac{1}{n-1}$ randomly and change the value $a_{ij}^{(m-1)}$ and $a_{ji}^{(m-1)}$ in matrix $\mathbf{A}^{(m-1)}$ to $1 - a_{ij}^{(m-1)}$ and $1 - a_{ji}^{(m-1)}$ and propose it as $\widetilde{\mathbf{A}}$. With a certain probability, we flip all elements in $\mathbf{A}^{(m-1)}$ from 0 to 1 (or 1 to 0) and propose it as $\widetilde{\mathbf{A}}$. Denote $\widetilde{\mathbf{M}} = \mathbf{I}_n + \widetilde{b}\mathbf{B} - \widetilde{\rho}(\widetilde{\mathbf{A}} \circ \mathbf{F})$ and calculate $\widetilde{\mathbf{q}}^* = \widetilde{\mathbf{M}}^{-1}\left(\mathbf{X}\delta^{(t-1)}\right)$. Simulate $\widetilde{\mathbf{q}}$ from $\mathscr{N}_n\left(\widetilde{\mathbf{q}}^*, \widetilde{\mathbf{M}}^{-1}\right)$. Then, with the acceptance probability

$$\widetilde{\alpha}(\widetilde{G}|G^{(m-1)}) = \min\left\{\frac{\exp\left(\Phi(\widetilde{G}, \widetilde{\mathbf{q}})\right)}{\exp\left(\Phi(G^{(m-1)}, \mathbf{q}^{(m-1)})\right)}, 1\right\},$$

[22] This step mimics the MH sampler in Snijders [2002] and Mele [2017]. In practice, we set $M = n^2$ where n is the size of the network.

set $G^{(m)}$ to $\widetilde{G}$ and $\mathbf{q}^{(m)} = \widetilde{\mathbf{q}}$, otherwise, set $G^{(m)} = G^{(m-1)}$ and $\mathbf{q}^{(m)} = \mathbf{q}^{(m-1)}$. After the M runs, collect $G'^{(M)}$ and $\mathbf{q}' = \mathbf{q}^{(M)}$.

(c) With the acceptance probability equals to

$$\begin{aligned}
&\alpha(\widetilde{\rho}, \widetilde{b} | \rho^{(t-1)}, b^{(t-1)}, G', \mathbf{q}') \\
&= \min \left\{ \frac{\mu\left(\mathbf{q}, G, \Upsilon^{(t-1)}, \widetilde{\rho}, \widetilde{b}\right)}{\mu\left(\mathbf{q}, G, \Upsilon^{(t-1)}, \rho^{(t-1)}, b^{(t-1)}\right)} \cdot \frac{\mu\left(\mathbf{q}', G'^{(t-1)}, \rho^{(t-1)}, b^{(t-1)}\right)}{\mu\left(\mathbf{q}', G'^{(t-1)}, \widetilde{\rho}, \widetilde{b}\right)} \cdot \frac{I((\widetilde{\rho}, \widetilde{b}) \in O)}{I((\rho^{(t-1)}, b^{(t-1)}) \in O)}, 1 \right\} \\
&= \min \left\{ \frac{\exp\left(\Phi(\mathbf{q}, G, \Upsilon^{(t-1)}, \widetilde{\rho}, \widetilde{b})\right)}{\exp\left(\Phi(\mathbf{q}, G, \Upsilon^{(t-1)}, \rho^{(t-1)}, b^{(t-1)})\right)} \cdot \frac{\exp\left(\Phi(\mathbf{q}', G'^{(t-1)}, \rho^{(t-1)}, b^{(t-1)})\right)}{\exp\left(\Phi(\mathbf{q}', G'^{(t-1)}, \widetilde{\rho}, \widetilde{b})\right)} \cdot \frac{I((\widetilde{\rho}, \widetilde{b}) \in O)}{I((\rho^{(t-1)}, b^{(t-1)}) \in O)}, 1 \right\},
\end{aligned}$$

set $(\rho^{(t)}, b^{(t)}) = (\widetilde{\rho}, \widetilde{b})$. Otherwise, set $(\rho^{(t)}, b^{(t)}) = (\rho^{(t-1)}, b^{(t-1)})$.

Step II. Simulate $\boldsymbol{\delta}^{(t)}$ from

$$P(\boldsymbol{\delta} | \mathbf{q}, G, \rho^{(t)}, b^{(t)}, \Upsilon^{(t-1)}) \propto \phi_l(\boldsymbol{\delta}; \boldsymbol{\delta}_0, \Delta_0) \cdot \mu(\mathbf{q} | G, \rho^{(t)}, b^{(t)}, \boldsymbol{\delta}, \Upsilon^{(t-1)}).$$

(a) propose $\widetilde{\boldsymbol{\delta}}$ from a random walk proposal density $T_2(\boldsymbol{\delta} | \boldsymbol{\delta}^{(t-1)})$.

(b) simulate an auxiliary data $\mathbf{q}'$ and G' by M runs of the MH algorithm based on

$$\mu(\mathbf{q}, G | \widetilde{\boldsymbol{\delta}}, \rho^{(t)}, b^{(t)}, \Upsilon^{(t-1)}) = \frac{\exp\left(\Phi\left(\mathbf{q}, G, \widetilde{\boldsymbol{\delta}}, \rho^{(t)}, b^{(t)}, \Upsilon^{(t-1)}\right)\right)}{\sum_{\mathfrak{g} \in \mathcal{G}^n} \int_{\mathcal{Q}^n} \exp\left(\Phi\left(\mathfrak{q}, \mathfrak{g}, \widetilde{\boldsymbol{\delta}}, \rho^{(t)}, b^{(t)}, \Upsilon^{(t-1)}\right)\right) d\mathfrak{q}}$$

starting from the observed network G.

(c) With the acceptance probability equals to

$$\begin{aligned}
&\alpha(\widetilde{\boldsymbol{\delta}} | \boldsymbol{\delta}^{(t-1)}, G', \mathbf{q}') \\
&= \min \left\{ \frac{\mu\left(\mathbf{q}, G, \widetilde{\boldsymbol{\delta}}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)}\right)}{\mu\left(\mathbf{q}, G, \boldsymbol{\delta}^{(t-1)}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)}\right)} \cdot \frac{\mu\left(\mathbf{q}', G', \boldsymbol{\delta}^{(t-1)}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)}\right)}{\mu\left(\mathbf{q}', G', \widetilde{\boldsymbol{\delta}}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)}\right)} \cdot \frac{\phi_l(\widetilde{\boldsymbol{\delta}}; \boldsymbol{\delta}_0, \Delta_0)}{\phi_l(\boldsymbol{\delta}^{(t-1)}; \boldsymbol{\delta}_0, \Delta_0)}, 1 \right\} \\
&= \min \left\{ \frac{\exp\left(\Phi(\mathbf{q}, G, \widetilde{\boldsymbol{\delta}}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)})\right)}{\exp\left(\Phi(\mathbf{q}, G, \boldsymbol{\delta}^{(t-1)}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)})\right)} \cdot \frac{\exp\left(\Phi(\mathbf{q}', G', \boldsymbol{\delta}^{(t-1)}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)})\right)}{\exp\left(\Phi(\mathbf{q}', G', \widetilde{\boldsymbol{\delta}}, \Upsilon^{(t-1)}, \rho^{(t)}, b^{(t)})\right)} \cdot \frac{\phi_l(\widetilde{\boldsymbol{\delta}}; \boldsymbol{\delta}_0, \Delta_0)}{\phi_l(\boldsymbol{\delta}^{(t-1)}; \boldsymbol{\delta}_0, \Delta_0)}, 1 \right\},
\end{aligned}$$

set $\boldsymbol{\delta}^{(t)} = \widetilde{\boldsymbol{\delta}}$. Otherwise, set $\boldsymbol{\delta}^{(t)} = \boldsymbol{\delta}^{(t-1)}$.

Step III. Simulate $\boldsymbol{\gamma}^{(t)}$ from $P\left(\boldsymbol{\gamma} \left| \mathbf{q}, G, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)} \right.\right)$ by the DMH algorithm.

(a) propose $\widetilde{\boldsymbol{\gamma}}$ from a random walk proposal density $T_3(\boldsymbol{\gamma} | \boldsymbol{\gamma}^{(t-1)})$

(b) simulate an auxiliary data $\mathbf{q}'$ and G' by M runs of the MH algorithm based on

$$\mu(\mathbf{q}, G | \widetilde{\boldsymbol{\gamma}}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}) = \frac{\exp\left(\Phi\left(\mathbf{q}, G, \widetilde{\boldsymbol{\gamma}}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)\right)}{\sum_{\mathfrak{g} \in \mathcal{G}^n} \int_{\mathcal{Q}^n} \exp\left(\Phi\left(\mathfrak{q}, \mathfrak{g}, \widetilde{\boldsymbol{\gamma}}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)\right) d\mathfrak{q}}$$

starting from the observed network G.

(d) With the acceptance probability equal to

$$\begin{aligned}
&\alpha(\widetilde{\boldsymbol{\gamma}}|\boldsymbol{\gamma}^{(t-1)}, G', \mathbf{q}') \\
&= \min\left\{\frac{\mu\left(G, \mathbf{q}, \widetilde{\boldsymbol{\gamma}}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)}{\mu\left(G, \mathbf{q}, \boldsymbol{\gamma}^{(t-1)}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)} \cdot \frac{\mu\left(G', \mathbf{q}', \boldsymbol{\gamma}^{(t-1)}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)}{\mu\left(G', \mathbf{q}', \widetilde{\boldsymbol{\gamma}}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)} \cdot \frac{\phi_r(\widetilde{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_0, \Gamma_0)}{\phi_r(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{\gamma}_0, \Gamma_0)}, 1\right\} \\
&= \min\left\{\frac{\exp\left(\Phi(G, \mathbf{q}, \widetilde{\boldsymbol{\gamma}}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)}{\exp\left(\Phi(G, \mathbf{q}, \boldsymbol{\gamma}^{(t-1)}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)} \cdot \frac{\exp\left(\Phi(G', \mathbf{q}', \boldsymbol{\gamma}^{(t-1)}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)}{\exp\left(\Phi(G', \mathbf{q}', \widetilde{\boldsymbol{\gamma}}, \Upsilon^{(t-1)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)} \cdot \frac{\phi_r(\widetilde{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_0, \Gamma_0)}{\phi_r(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{\gamma}_0, \Gamma_0)}, 1\right\},
\end{aligned}$$

set $\boldsymbol{\gamma}^{(t)} = \widetilde{\boldsymbol{\gamma}}$. Otherwise, set $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)}$.

Step IV. Simulate $\varkappa^{(t)}$ from $P\left(\varkappa \left| \mathbf{q}, G, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right.\right)$ by the DMH algorithm.

(a) propose $\widetilde{\varkappa}$ from a random walk proposal density $T_4(\varkappa|\varkappa^{(t-1)})$

(b) simulate an auxiliary data $\mathbf{q}'$ and G' by M runs of the MH algorithm based on

$$\mu(\mathbf{q}, G|\widetilde{\varkappa}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}) = \frac{\exp\left(\Phi\left(\mathbf{q}, G, \widetilde{\varkappa}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)\right)}{\sum_{\mathfrak{g}\in\mathcal{G}^n}\int_{\mathcal{Q}^n}\exp\left(\Phi\left(\mathfrak{q}, \mathfrak{g}, \widetilde{\varkappa}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)\right)d\mathfrak{q}}$$

starting from the observed network G.

(d) With the acceptance probability equal to

$$\begin{aligned}
&\alpha(\widetilde{\varkappa}|\varkappa^{(t-1)}, G', \mathbf{q}') \\
&= \min\left\{\frac{\mu\left(G, \mathbf{q}, \widetilde{\varkappa}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)}{\mu\left(G, \mathbf{q}, \varkappa^{(t-1)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)} \cdot \frac{\mu\left(G', \mathbf{q}'^{(t-1)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)}{\mu\left(G', \mathbf{q}', \widetilde{\varkappa}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)}\right)} \cdot \frac{\phi_r(\widetilde{\varkappa}|\varkappa_0, \mathbf{K}_0)}{\phi_r(\varkappa^{(t-1)}|\varkappa_0, \mathbf{K}_0)}, 1\right\} \\
&= \min\left\{\frac{\exp\left(\Phi(G, \mathbf{q}, \widetilde{\varkappa}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)}{\exp\left(\Phi(G, \mathbf{q}, \varkappa^{(t-1)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)} \cdot \frac{\exp\left(\Phi(G', \mathbf{q}'^{(t-1)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)}{\exp\left(\Phi(G', \mathbf{q}', \widetilde{\varkappa}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}, \rho^{(t)}, b^{(t)})\right)} \cdot \frac{\phi_r(\widetilde{\varkappa}|\varkappa_0, \mathbf{K}_0)}{\phi_r(\varkappa^{(t-1)}|\varkappa_0, \mathbf{K}_0)}, 1\right\},
\end{aligned}$$

set $\boldsymbol{\gamma}^{(t)} = \widetilde{\boldsymbol{\gamma}}$. Otherwise, set $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)}$.

G.2. Implementation of the AEX Algorithm

The AEX algorithm consists of two Markov chains running in parallel. Let the subscript t denote the t^{th} iteration of two chains. In the first chain, we simulate auxiliary network and output sample $(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)})$ from $\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\widetilde{\boldsymbol{\theta}}^{(t)})$, where $\widetilde{\boldsymbol{\theta}}^{(t)}$ is sampled from the set of pre-specified parameter points $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m)$. In our applications, we set $m = 50$ and $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m)$ are chosen from the DMH draws using the Max-Min procedure suggested in Liang et al. [2015]. The index function $H(\widetilde{\boldsymbol{\theta}}^{(t)})$ denotes the order index of $\widetilde{\boldsymbol{\theta}}^{(t)}$, i.e., $H(\widetilde{\boldsymbol{\theta}}^{(t)}) = i$ if $\widetilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}_i$. Let $\boldsymbol{p} = (p_1, \cdots, p_m)$ be the desired sampling frequencies from the respective distributions $\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}_1), \cdots, \mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}_m)$, where $0 < p_i < 1$ and $\sum_{i=1}^m p_i = 1$. We follow Liang et al. [2015] to set $p_1 = \cdots = p_m = \frac{1}{m}$. As required by the SAMC algorithm, we specify a *gain factor* sequence, $\{a_t\}$, which is a positive, nonincreasing sequence satisfying the following condition:

(A_1) $\lim_{t\to\infty} a_t = 0$, $\quad \sum_{t=1}^{\infty} a_t = \infty$, $\quad \sum_{t=1}^{\infty} a_t^k < \infty$ for some $k \in (1, 2]$.

Following Liang et al. [2015], we set $a_t = \frac{t_0}{\max(t_0, t)}$ with a known constant $t_0 > 1$. Since a larger value of t_0 will help the auxiliary chain to reach each distribution $\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}_i)$ more quickly, we

set $t_0 = 20,000$ in this paper. Also, we let $w_i^{(t)}$ denote an *abundance factor* attached to each distribution $\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}_i)$ at iteration t and $w^{(t)} = (w_1^{(t)}, \cdots, w_m^{(t)})$. We set the initial values $w_1^{(0)} = \cdots = w_m^{(0)} = 1$ in the simulation. Finally, let $\Omega^{(t)}$= $(\widetilde{G}^{(1)}, \cdots, \widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(1)}, \cdots, \widetilde{\mathbf{q}}^{(t)}, \widetilde{\boldsymbol{\theta}}^{(1)}, \cdots, \widetilde{\boldsymbol{\theta}}^{(t)}, H(\widetilde{\boldsymbol{\theta}}^{(1)}), \cdots, H(\widetilde{\boldsymbol{\theta}}^{(t)}), w^{(1)}, \cdots, w^{(t)})$ denote the information in the auxiliary chain that we collect up to iteration t. In the second chain, we draw $\boldsymbol{\theta}^{(t)}$ by the exchange algorithm from the target posterior distribution $\mu(\boldsymbol{\theta}|\mathbf{q}, G)$. Explicitly, the AEX algorithm is implemented by Part I and Part II at each t^{th} iteration described as follows:

Part I. Auxiliary network simulation

1. Choose to update $\widetilde{\boldsymbol{\theta}}^{(t)}$ or $\widetilde{G}^{(t)}$ with a pre-specified probability. In our application, we choose 75% for updating $\widetilde{\boldsymbol{\theta}}^{(t)}$ and 25% for updating $\widetilde{G}^{(t)}$.
   (a) Update $\widetilde{\boldsymbol{\theta}}^{(t)}$: Select $\widetilde{\boldsymbol{\theta}}'$ from the set $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m)$ according to a proposal distribution $T_1(\cdot|\widetilde{\boldsymbol{\theta}}^{(t-1)})$. With the probability

$$\widetilde{\alpha}\left(\widetilde{\boldsymbol{\theta}}' \,\middle|\, \widetilde{\boldsymbol{\theta}}^{(t-1)}\right) = \min\left\{ \frac{w^{(t-1)}_{H(\widetilde{\boldsymbol{\theta}}^{(t-1)})} \exp(\Phi(\widetilde{G}^{(t-1)}, \widetilde{\mathbf{q}}^{(t-1)}, \widetilde{\boldsymbol{\theta}}')) T_1(\widetilde{\boldsymbol{\theta}}^{(t-1)}|\widetilde{\boldsymbol{\theta}}')}{w^{(t-1)}_{H(\widetilde{\boldsymbol{\theta}}')} \exp(\Phi(\widetilde{G}^{(t-1)}, \widetilde{\mathbf{q}}^{(t-1)}, \widetilde{\boldsymbol{\theta}}^{(t-1)})) T_1(\widetilde{\boldsymbol{\theta}}'|\widetilde{\boldsymbol{\theta}}^{(t-1)})}, 1 \right\}, \tag{G.23}$$

   update $(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}) = (\widetilde{\boldsymbol{\theta}}', \widetilde{G}^{(t-1)}, \widetilde{\mathbf{q}}^{(t-1)})$. Otherwise,
   set $(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}) = (\widetilde{\boldsymbol{\theta}}^{(t-1)}, \widetilde{G}^{(t-1)}, \widetilde{\mathbf{q}}^{(t-1)})$.
   (b) Update $(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)})$: simulate $\widetilde{G}^{(t)}$ and $\widetilde{\mathbf{q}}^{(t)}$ from $\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\widetilde{\boldsymbol{\theta}}^{(t-1)})$ via few MH updates starting from $(\widetilde{G}^{(t-1)}, \widetilde{\mathbf{q}}^{(t-1)})$. Then set $\widetilde{\boldsymbol{\theta}}^{(t)} = \widetilde{\boldsymbol{\theta}}^{(t-1)}$.
2. For $i = 1, \cdots, m$, update the abundance factor $w_i^{(t)}$ by

$$\log(w_i^{(t)}) = \log(w_i^{(t-1)}) + a_t(e_{i,t} - p_i),$$

   where $e_{i,t} = 1$ if $\widetilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}_i$ and 0 otherwise.
3. Append $(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, H(\widetilde{\boldsymbol{\theta}}^{(t)}), w^{(t)})$ to the collection $\Omega^{(t-1)}$ to form $\Omega^{(t)}$.

Part II. Adaptive Exchange algorithm for target parameter

4. Propose a candidate $\boldsymbol{\theta}'$ from a random walk proposal distribution $T_2(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})$.
5. Re-sample an auxiliary network and output data $(G', \mathbf{q}')$ from the collection $\Omega^{(t)}$ via an importance sampling procedure. With the probability

$$P\left((G', \mathbf{q}') = (\widetilde{G}^{(i)}, \widetilde{\mathbf{q}}^{(i)})\right) = \frac{\sum_{j=1}^{t} \left\{ w^{(j)}_{H(\widetilde{\boldsymbol{\theta}}^{(j)})} \frac{\exp(\Phi(\widetilde{G}^{(j)}, \widetilde{\mathbf{q}}^{(j)}, \boldsymbol{\theta}'))}{\exp(\Phi(\widetilde{G}^{(j)}, \widetilde{\mathbf{q}}^{(j)}, \widetilde{\boldsymbol{\theta}}^{(j)}))} I\left(\widetilde{G}^{(j)} = \widetilde{G}^{(i)}\right) \right\}}{\sum_{j=1}^{t} \left\{ w^{(j)}_{H(\widetilde{\boldsymbol{\theta}}^{(j)})} \frac{\exp(\Phi(\widetilde{G}_j, \widetilde{\mathbf{q}}^{(j)}, \boldsymbol{\theta}'))}{\exp(\Phi(\widetilde{G}^{(j)}, \widetilde{\mathbf{q}}^{(j)}, \widetilde{\boldsymbol{\theta}}^{(j)}))} \right\}}, \tag{G.24}$$

   choose $(G', \mathbf{q}') = (\widetilde{G}^{(i)}, \widetilde{\mathbf{q}}^{(i)})$, $i \in (1, 2, \cdots, t)$.
6. Implement the exchange algorithm. With the probability

$$\begin{aligned} &\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)}, G', \mathbf{q}') \\ &= \min\left\{ \frac{\exp(\Phi(G, \mathbf{q}, \boldsymbol{\theta}'))}{\exp(\Phi(G, \mathbf{q}, \boldsymbol{\theta}^{(t-1)}))} \cdot \frac{\exp(\Phi(G', \mathbf{q}', \boldsymbol{\theta}^{(t-1)}))}{\exp(\Phi(G', \mathbf{q}', \boldsymbol{\theta}'))} \cdot \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)})}, 1 \right\}, \end{aligned} \tag{G.25}$$

   set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}'$. Otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

G.3. Convergence of the AEX Algorithm

In this appendix, we outline the ideas and main results behind the convergence proof of the AEX algorithm. Interested readers are referred to Jin et al. [2013] and the supplementary material of Liang et al. [2015] for the completed version of proof. From Part II of the AEX algorithm, one can see that auxiliary sample $(G', \mathbf{q}')$ is drawn via a dynamic importance sampling procedure in Equation (G.24), which implies that the underlying proposal distribution for auxiliary networks changes from iteration to iteration. Therefore, AEX falls into the class of adaptive MCMC algorithms with varying stationary distributions and requires an unconventional convergence theory [Fort et al., 2011; Roberts and Rosenthal, 2007].

Liang et al. [2015] provided an ergodicity theorem for adaptive Markov chains with varying stationary distributions, which can be used to prove ergodicity of the AEX algorithm. Consider a state space, $(\mathbb{X}, \mathcal{H})$, where $X_t \in \mathbb{X}$ denotes the state of the Markov chain at iteration t and $\mathcal{H}$ denotes the Borel set defined on $\mathbb{X}$. Let γ_t denote realization of the adoption index $\Gamma_t \in \mathbb{Y}$ and $\mathcal{F}_t = \sigma(X_0, \cdots, X_t, \Gamma_0, \cdots, \Gamma_t)$ be the filtration generated by $\{X_i, \Gamma_i\}_{i=0}^t$. Let P_{γ_t} denote the transition kernel at iteration t and thus,

$$P(X_{t+1} \in B | X_t = x, \Gamma_t = \gamma_t, \mathcal{F}_{t-1}) = P_{\gamma_t}(x, B), \quad x \in \mathbb{X}, \gamma_t \in \mathbb{Y}, B \in \mathcal{H}.$$

Let $P_\gamma^h(x, B) = P_\gamma(X_h \in B | X_0 = x)$ denote the h-step transition probability for the Markov chain with the fixed transition kernel P_γ and the initial condition $X_0 = x$. Also let $P^h((x, \gamma), B) = P(X_h \in B | X_0 = x, \Gamma_0 = \gamma, \mathcal{F}_{h-1})$, $B \in \mathcal{H}$, denote the h-step transition probability for the adaptive Markov chain with the initial conditions $X_0 = x$ and $\Gamma_0 = \gamma$. An adaptive Markov chain is called ergodic if $\lim_{t\to\infty} V(x, \gamma, t) = 0$ for all $x \in \mathbb{X}$ and $\gamma \in \mathbb{Y}$, where

$$V(x, \gamma, t) = \| P^t((x, \gamma), \cdot) - \pi(\cdot) \| = \sup_{B \in \mathcal{H}} \| P^t((x, \gamma), B) - \pi(B) \|$$

denotes the total variation distance between the distribution of the adaptive chain at iteration t and the target distribution $\pi(\cdot)$.

Theorem 2. *Consider an adaptive Markov chain defined on the state space $(\mathbb{X}, \mathcal{H})$ with the adoption index $\Gamma_t \in \mathbb{Y}$. The adoptive Markov chain is ergodic if the following conditions are satisfied:*

(a) (Stationarity) There exist a stationary distribution $\pi_{\gamma_t}(\cdot)$ for each transition kernel P_{γ_t}.

(b) (Asymptotic Simultaneous Uniform Ergodicity) For any $u > 0$, there exist a measurable set $E(u)$ in the probability space such that $Pr(E(u)) \geq 1 - u$ and on this set $E(u)$, for any $\epsilon > 0$, there exist $t(\epsilon) \in \mathbb{N}$ and $n(\epsilon) \in \mathbb{N}$ such that

$$\sup_{x \in \mathbb{X}} \| P_{\gamma_t}^n(x, \cdot) - \pi(\cdot) \| \leq \epsilon,$$

for all $t > t(\epsilon)$ and $n > n(\epsilon)$.

(c) (Diminishing Adoption) let $D_t = \sup_{x \in \mathbb{X}} \| P_{\Gamma_{t+1}}(x, \cdot) - P_{\Gamma_t}(x, \cdot) \|$ and $\lim_{t\to\infty} D_t = 0$ in probability.

From conditions (a) and (b) of Theorem 2, it is implied that on the set $E(u)$, for any ϵ,

$$\| \pi_{\gamma_t}(\cdot) - \pi(\cdot) \| = \| \pi_{\gamma_t}(\cdot) P_{\gamma_t}^h - \pi(\cdot) \| \leq \sup_{x \in \mathbb{X}} \| P_{\gamma_t}^n(x, \cdot) - \pi(\cdot) \| \leq \epsilon,$$

for $t > t(\epsilon)$ and $n > n(\epsilon)$. Furthermore, by the triangular inequality,

$$\| P^n_{\gamma_t}(x,\cdot) - \pi_{\gamma_t}(\cdot) \| < \| P^n_{\gamma_t}(x,\cdot) - \pi(\cdot) \| + \| \pi_{\gamma_t}(\cdot) - \pi(\cdot) \| < 2\epsilon,$$

for $t > t(\epsilon)$ and $n > n(\epsilon)$. Thus, the simultaneous uniform ergodicity of $\{P_{\gamma_t}\}$ on the set $E(u)$ is established. Given condition (c) of Theorem 2 and the simultaneous uniform ergodicity of $\{P_{\gamma_t}\}$, Liang et al. [2015] show the weak law of large numbers for an adaptive Markov chain in the following theorem:

Theorem 3. *Consider an adaptive Markov chain defined on the state space $(\mathbb{X}, \mathcal{H})$. Let $g(\cdot)$ be a bounded measurable function. Suppose that conditions (a), (b), and (c) of Theorem 2 hold, then $\frac{1}{T}\sum_{t=1}^{T} g(X_t) \to \pi(g)$, in probability as $T \to \infty$, where $\pi(g) = \int_{\mathbb{X}} g(x)\pi(dx)$.*

The rest of procedure is to show the AEX algorithm satisfies three conditions of Theorem 2 and thus it is ergodic and the weak law of large number holds for the average of sample path $\{\boldsymbol{\theta}^{(t)}\}$. To begin with, we assume that in the auxiliary chain, the Markov transition kernel, P_w, for updating the state variable $X_t = (\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \widetilde{\boldsymbol{\theta}}^{(t)}) \in \mathbb{X}$ via the SAMC algorithm satisfies the following Doeblin condition:

(A_2) For any given $w \in \varpi$, the Markov transition kernel P_w is irreducible and aperiodic. In addition, there exist an integer h, a real number $0 < \delta < 1$, and a probability measure ν such that for any compact subset $\mathfrak{K} \subset \varpi$,

$$\inf_{w \in \mathfrak{K}} P^h_w(x, B) \geq \delta\nu(B), \quad \forall x \in \mathbb{X},\ \forall B \in \mathcal{H},$$

where $\mathcal{H}$ denotes the Borel set of $\mathbb{X}$.

The condition (A_2) will be assured by Theorem 2.2 of Roberts and Tweedie [1996] if $\mathbb{X}$ is compact, the potential function $\Phi(G', \epsilon(\boldsymbol{\theta}), \boldsymbol{\theta})$ is bounded away from 0 and ∞ on $\mathbb{X}$, and the proposal distribution $T(y|x)$ satisfies the local positive condition:

(local positive condition) For every $x \in \mathbb{X}$, there exist $\epsilon_1 > 0$, and $\epsilon_2 > 0$ such that $|x - y| \leq \epsilon_1 \Rightarrow T(y|x) \geq \epsilon_2$.

The following Lemma 2 shows that draws of $(G', \mathbf{q}')$ from the dynamic importance sampler of Equation (G.24) converges to the distribution of $\mu(\cdot|\boldsymbol{\theta}')$ almost surely when the number of iterations goes to infinity.

Lemma 2. *Assume (i) conditions (A_1) and (A_2) are satisfied; (ii) $\mathbb{X}$ is compact; (iii) $\exp(\Psi(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}))$ is bounded away from 0 and ∞ on $\mathbb{X} \times \Theta$; Given a total of N iterations, let $\{\mathfrak{g}^{(k)}, \mathfrak{q}^{(k)}\}_{k=1}^{n}$ be n distinct values of $(\widetilde{G}, \widetilde{q})$. Re-sample a random sample $(G', \mathbf{q}')$ from $\Omega^{(N)}$ such that*

$$P((G', \mathbf{q}') = (\mathfrak{g}^{(k)}, \mathfrak{q}^{(k)})|\boldsymbol{\theta}') = \frac{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_i^{(t)} \frac{\exp(\Psi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}'))}{\exp(\Psi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}_i))} I\left(H(\widetilde{\boldsymbol{\theta}}^{(t)}) = i \text{ and } (\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}) = (\mathfrak{g}^{(k)}, \mathfrak{q}^{(k)})\right)\right\}}{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_i^{(t)} \frac{\exp(\Psi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}'))}{\exp(\Psi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}_i))} I\left(H(\widetilde{\boldsymbol{\theta}}^{(t)}) = i\right)\right\}}, \tag{G.26}$$

for $k = 1, \cdots, n$, then the distribution of $(G', \mathbf{q}')$ converges to $\mu(G', \mathbf{q}'|\boldsymbol{\theta}')$ almost surely as $N \to \infty$.

Proof of Lemma 2. By the assumption that $\mathbb{X}$ is compact and $\exp(\Psi(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}))$ is bounded away from 0 and ∞, it follows from the convergence and the strong law of large numbers of SAMC [Liang et al., 2015] that

$$\begin{aligned}
&\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_i^{(t)}\frac{\exp(\Psi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}'))}{\exp(\Psi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}_i))}I\left(H(\widetilde{\boldsymbol{\theta}}^{(t)}) = i\right)\right\} \\
&\quad \to \sum_{i=1}^{m}\int_{\mathbb{X}}\frac{\kappa(\boldsymbol{\theta}_i)}{p_i}\frac{\exp(\Psi(\widetilde{G}, \widetilde{\mathbf{q}}, \boldsymbol{\theta}'))}{\exp(\Psi(\widetilde{G}, \widetilde{\mathbf{q}}, \boldsymbol{\theta}_i))}p_i\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}_i)d\widetilde{G}d\widetilde{\mathbf{q}} \\
&\quad = m\kappa(\boldsymbol{\theta}'), \quad a.s.,
\end{aligned} \tag{G.27}$$

where $\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}_i) = \frac{\exp(\Phi(\widetilde{G}, \widetilde{\mathbf{q}}, \boldsymbol{\theta}_i))}{\kappa(\boldsymbol{\theta}_i)}$ and $\kappa(\boldsymbol{\theta}_i) = \int_{\mathbb{X}}\exp(\Phi(\widetilde{G}, \widetilde{\mathbf{q}}, \boldsymbol{\theta}_i))$. Similarly, for any Borel set $A \in \mathbb{X}$,

$$\begin{aligned}
&\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_i^{(t)}\frac{\exp(\Phi(\widetilde{G}_t, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}'))}{\exp(\Phi(\widetilde{G}_t, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}_i))}I\left(H(\widetilde{\boldsymbol{\theta}}^{(t)}) = i \text{ and } (\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}) \in A\right)\right\} \\
&\quad \to \sum_{i=1}^{m}\int_{A}\frac{\kappa(\boldsymbol{\theta}_i)}{p_i}\frac{\exp(\Phi(\widetilde{G}, \widetilde{\mathbf{q}}, \boldsymbol{\theta}'))}{\exp(\Phi(\widetilde{G}, \widetilde{\mathbf{q}}, \boldsymbol{\theta}_i))}p_i\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}_i)d\widetilde{G}d\widetilde{\mathbf{q}} \\
&\quad = m\int_{A}\exp(\Phi(\widetilde{G}, \widetilde{\mathbf{q}}, \boldsymbol{\theta}'))d\widetilde{G}d\widetilde{\mathbf{q}}.
\end{aligned} \tag{G.28}$$

Putting Equations (G.27) and (G.28) together, as $N \to \infty$, we have

$$\begin{aligned}
&P((G', \mathbf{q}') \in A|\Omega^{(N)}, \boldsymbol{\theta}') \qquad (G.29) \\
&= \frac{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_i^{(t)}\frac{\exp(\Phi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}'))}{\exp(\Phi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}_i))}I\left(H(\widetilde{\boldsymbol{\theta}}^{(t)}) = i \text{ and } (\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}) \in A\right)\right\}}{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_i^{(t)}\frac{\exp(\Phi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}'))}{\exp(\Phi(\widetilde{G}^{(t)}, \widetilde{\mathbf{q}}^{(t)}, \boldsymbol{\theta}_i))}I\left(H(\widetilde{\boldsymbol{\theta}}^{(t)}) = i\right)\right\}} \\
&\to \int_{A}\mu(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}')d\widetilde{G}d\widetilde{\mathbf{q}}, \quad a.s. \qquad (G.30)
\end{aligned}$$

Finally, by Lebesque's dominated convergence theorem,

$$P((G', \mathbf{q}') \in A|\theta') = E[P((G', \mathbf{q}') \in A|\Omega^{(N)}; \boldsymbol{\theta}')] \to \int_{A} f(\widetilde{G}, \widetilde{\mathbf{q}}|\boldsymbol{\theta}')d\widetilde{G}d\widetilde{\mathbf{q}}. \tag{G.31}$$

□

Notice that $\{\boldsymbol{\theta}^{(t)}\}$ from the AEX algorithm form an adaptive Markov chain with the transition kernel given by

$$\begin{aligned}
&\widetilde{P}_l(\boldsymbol{\theta}, d\boldsymbol{\theta}') = \\
&\int_{\mathbb{X}}\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}, G', \mathbf{q}')T_1(d\boldsymbol{\theta}'|\boldsymbol{\theta})\nu_l(G', \mathbf{q}'|\boldsymbol{\theta}')dG'd\mathbf{q}' + \delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}')\left[1 - \int_{\Theta\times\mathbb{X}}\alpha(\boldsymbol{\vartheta}'|\boldsymbol{\theta}, G', \mathbf{q}')T_1(d\boldsymbol{\vartheta}'|\boldsymbol{\theta})\nu_l(G', \mathbf{q}'|\boldsymbol{\vartheta}')dG'd\mathbf{q}'\right],
\end{aligned}$$

where $\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}, G', \mathbf{q}')$ is defined in Equation (G.25), $\delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}') = 1$ if $\boldsymbol{\theta} \in d\boldsymbol{\theta}'$ and 0 otherwise, $T_1(d\boldsymbol{\theta}'|\boldsymbol{\theta})$ is the proposal distribution, l denotes the cardinality of the set $\Omega^{(l)}$, i.e., $l = |\Omega^{(l)}|$, and $\nu_l(G', \mathbf{q}'|\boldsymbol{\theta}')$ denotes the distribution of $(G', \mathbf{q}')$ re-sampled from $\Omega^{(l)}$. Different from the transition kernel of AEX, the transition kernel of the exchange algorithm depends on the perfect draw of

$(G', \mathbf{q}')$ from $\mu(\cdot|\boldsymbol{\vartheta}')$, which is given by

$$P(\boldsymbol{\theta}, d\boldsymbol{\theta}') =$$
$$\int_{\mathbb{X}} \alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}, G', \mathbf{q}') T_1(d\boldsymbol{\theta}'|\boldsymbol{\theta}) \mu(G', \mathbf{q}'|\boldsymbol{\theta}') dG' d\mathbf{q}' + \delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}') \left[1 - \int_{\Theta \times \mathbb{X}} \alpha(\boldsymbol{\vartheta}'|\boldsymbol{\theta}, G', \mathbf{q}') T_1(d\boldsymbol{\vartheta}'|\boldsymbol{\theta}) \mu(G', \mathbf{q}'|\boldsymbol{\vartheta}') dG' d\mathbf{q}'\right],$$

where $\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}, G', \mathbf{q}')$ is defined in Equation (31)

Lemma 3. *Assume that (i) conditions (A_1) and (A_2) are satisfied; (ii) both Θ and $\mathbb{X}$ are compact; (iii) the potential function $\Phi(G', \epsilon(\boldsymbol{\theta}), \boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$ for all $G' \in \mathbb{X}$ and bounded away from 0 and ∞ on $\mathbb{X} \times \Theta$; (iv) $\pi(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ are continuously differentiable in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Define $\widetilde{D}_l = \sup_{\boldsymbol{\theta} \in \Theta} \parallel \widetilde{P}_l(\boldsymbol{\theta}, \cdot) - P(\boldsymbol{\theta}, \cdot) \parallel$, then $\widetilde{D}_l \to 0$ almost surely as $l \to \infty$.*

Define $D_l = \sup_{\boldsymbol{\theta} \in \Theta} \parallel \widetilde{P}_{l+1}(\boldsymbol{\theta}, \cdot) - \widetilde{P}_l(\boldsymbol{\theta}, \cdot) \parallel$. Since $D_l \leq \sup_{\boldsymbol{\theta} \in \Theta} \parallel \widetilde{P}_{l+1}(\boldsymbol{\theta}, \cdot) - P(\boldsymbol{\theta}, \cdot) \parallel + \sup_{\boldsymbol{\theta} \in \Theta} \parallel \widetilde{P}_l(\boldsymbol{\theta}, \cdot) - P(\boldsymbol{\theta}, \cdot) \parallel$, we have $\lim_{l \to \infty} D_l = 0$ almost surely by Lemma 3. Thus, $\widetilde{P}_l$ satisfies condition (c) of Theorem D1. It is known that the transition kernel of the exchange algorithm, $P(\boldsymbol{\theta}, d\boldsymbol{\theta}')$, is irreducible and aperiodic and admits an invariant limit distribution. Lemma 4 shows that the transition kernel of the AEX algorithm, $\widetilde{P}_l(\boldsymbol{\theta}, d\boldsymbol{\theta}')$, also has these properties and thus satisfies condition (a) of Theorem 2.

Lemma 4. *Assume (i) the conditions of Lemma 3 are satisfied; (ii) P is irreducible and aperiodic and admits an invariant distribution, then $\widetilde{P}_l$ is irreducible and aperiodic, and hence exists a stationary distribution $\pi_l(\boldsymbol{\theta}|G', \mathbf{q}')$ such that for any $\boldsymbol{\theta}_0 \in \Theta$,*

$$\lim_{k \to \infty} \parallel \widetilde{P}_l^k(\boldsymbol{\theta}_0, \cdot) - \pi_l(\cdot|G', \mathbf{q}') \parallel = 0.$$

Lemma 5 establishes the asymptotic simultaneous uniform ergodicity of the kernel $\widetilde{P}_l$'s.

Lemma 5. *Assume the conditions of Lemma 4 are satisfied. If the proposal $q(\cdot, \cdot)$ satisfies the local positive condition, then for any $e > 0$, there exist a measurable set $E(e)$ in the probability space such that $P(E(e)) > 1 - e$ and on this set $E(e)$, for any $\epsilon > 0$, there exist $n(\epsilon) \in \mathbb{N}$ and $l(\epsilon) \in \mathbb{N}$ such that for any $n > n(\epsilon)$ and $l > l(\epsilon)$,*

$$\parallel \widetilde{P}_l^n(\boldsymbol{\theta}_0, \cdot) - \pi(\cdot|G, \mathbf{q}) \parallel \leq \epsilon, \quad \text{for all } \boldsymbol{\theta}_0 \in \Theta.$$

Putting Lemmas 3, 4, and 5 altogether, we have the following theorem in regard to ergodicity and the weak law of large number for AEX.

Theorem 4. *Assume the conditions of Lemma 5 hold. If the proposal $T_1(\cdot, \cdot)$ satisfies the local positive condition and the unnormalized density function $\exp(\Phi(G', \mathbf{q}', \boldsymbol{\theta}))$ is upper semi-continuous in $\boldsymbol{\theta}$ for all $(G', \mathbf{q}') \in \mathbb{X}$, then the adaptive exchange algorithm is ergodic and for any bounded measurable function f,*

$$\frac{1}{T} \sum_{t=1}^{T} f(\boldsymbol{\theta}^{(t)}) \to \pi(f|G, \mathbf{q}), \quad \textit{in probability},$$

as $T \to \infty$, where $\pi(f|G, \mathbf{q}) = \int_{\Theta} f(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|G, \mathbf{q}) d\boldsymbol{\theta}$.

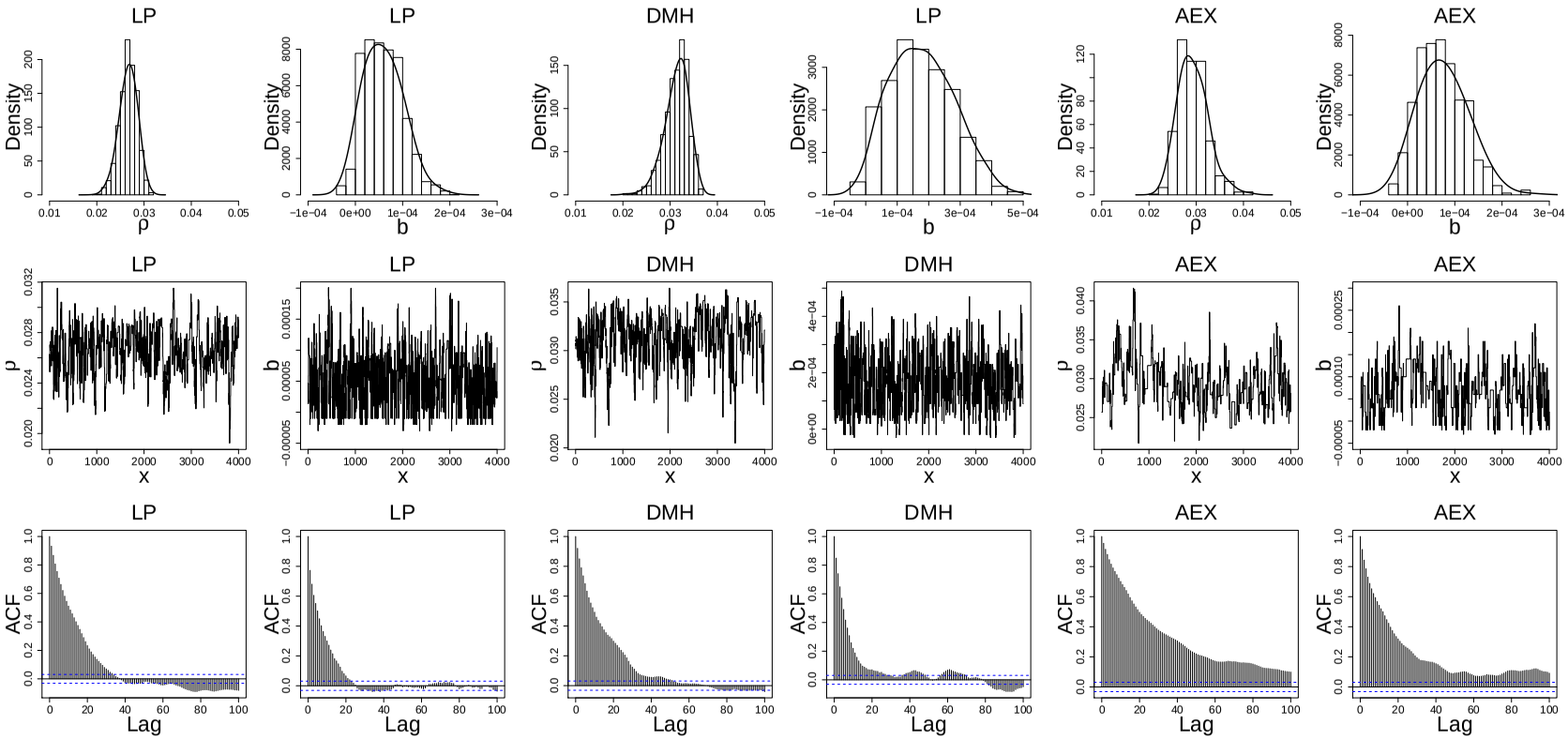


Figure G.1: Plots of MCMC draws for estimating ρ and b in SIC-28 subsample.

G.4. Fixed Network

When the network link update rates vanish (by setting the rates $\tau = \xi = 0$ of the stochastic process in Definition 1), i.e. the network is fixed, or just looking at a time interval between any two link changes where all firms adjust output levels instantaneously, we can study firms' output decisions conditional on a given fixed network configuration G. Assuming that the inverse noise parameter ϑ is large, output levels chosen by firms will be close to the potential maximizer $\mathbf{q}^*$. We can then make the following Taylor expansion for the potential function

$$\Phi(\mathbf{q}, G) = \Phi(\mathbf{q}^*, G) + (\mathbf{q} - \mathbf{q}^*)^\top \nabla\Phi(\mathbf{q}^*, G) + \frac{1}{2}(\mathbf{q} - \mathbf{q}^*)^\top \Delta\Phi(\mathbf{q}^*, G)(\mathbf{q} - \mathbf{q}^*) + o(\|\mathbf{q} - \mathbf{q}^*\|).$$

Noting that the gradient vanishes at $\mathbf{q}^*$, i.e., $\nabla\Phi(\mathbf{q}^*, G) = 0$, and dropping terms of the order $o(\|\mathbf{q} - \mathbf{q}^*\|)$ we can write

$$\Phi(\mathbf{q}, G) \approx \Phi(\mathbf{q}^*, G) + \frac{1}{2}(\mathbf{q} - \mathbf{q}^*)^\top \Delta\Phi(\mathbf{q}^*, G)(\mathbf{q} - \mathbf{q}^*).$$

Inserting the above equation into the Gibbs measure gives

$$\mu^\vartheta(\mathbf{q}, G) = \frac{1}{\mathscr{Z}_\vartheta} e^{\vartheta\Phi(G,\mathbf{q})} \approx \frac{1}{\mathscr{Z}_\vartheta} e^{\vartheta\Phi(\mathbf{q}^*,G)} \exp\left\{\frac{1}{2}\vartheta(\mathbf{q} - \mathbf{q}^*)^\top \Delta\Phi(\mathbf{q}^*, G)(\mathbf{q} - \mathbf{q}^*)\right\}.$$

The partition function can be written as

$$\begin{aligned}
\mathscr{Z}_\vartheta &= \int_{\mathcal{Q}^n} e^{\vartheta\Phi(\mathbf{q},G)} d\mathbf{q} \\
&\approx e^{\vartheta\Phi(\mathbf{q}^*,G)} \int_{\mathcal{Q}^n} \exp\left\{\frac{1}{2}\vartheta(\mathbf{q} - \mathbf{q}^*)^\top \Delta\Phi(\mathbf{q}^*, G)(\mathbf{q} - \mathbf{q}^*)\right\} d\mathbf{q} \\
&= e^{\vartheta\Phi(\mathbf{q}^*,G)} \left(\frac{2\pi}{\vartheta}\right)^{\frac{n}{2}} |-\nabla\Phi(\mathbf{q}^*, G)|^{-\frac{1}{2}}.
\end{aligned}$$

Therefore, the above Laplace approximation of the Gibbs measure yields [Wong, 2001, Theorem 3, p. 495]

$$\mu^\vartheta(\mathbf{q}, G) \approx \left(\frac{2\pi}{\vartheta}\right)^{-\frac{n}{2}} |-\nabla\Phi(\mathbf{q}^*, G)|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\vartheta(\mathbf{q} - \mathbf{q}^*)^\top(-\Delta\Phi(\mathbf{q}^*, G))(\mathbf{q} - \mathbf{q}^*)\right\}.$$

where the potential function $\Phi(\mathbf{q}, G)$ is given by Equation (25). The gradient can be written as $\nabla\Phi(\mathbf{q}, G) = \boldsymbol{\eta} - \mathbf{M}(G)\mathbf{q}$ and the Hessian is given by $\Delta\Phi(\mathbf{q}, G) = -\mathbf{M}(G)$. The FOCs are then given by $\nabla\Phi(\mathbf{q}^*, G) = \boldsymbol{\eta} - \mathbf{M}(G)\mathbf{q}^* = 0$ so that $\mathbf{q}^* = \mathbf{M}(G)^{-1}\mathbf{X}\boldsymbol{\delta}$ when the matrix $\mathbf{M}(G)$ is invertible and $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\delta}$.[23] Hence, the density of output levels $\mathbf{q}$ with the (exogenous) network G can be rewritten as

$$\mu^\vartheta(\mathbf{q}|G) \approx \left(\frac{2\pi}{\vartheta}\right)^{-\frac{n}{2}} |\mathbf{M}(G)|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{q} - \mathbf{M}(G)^{-1}\mathbf{X}\boldsymbol{\delta}\right)^\top \mathbf{M}(G)\left(\mathbf{q} - \mathbf{M}(G)^{-1}\mathbf{X}\boldsymbol{\delta}\right)\right\}, \quad \text{(G.32)}$$

which implies that the output $\mathbf{q}$, conditional on the R&D network G, follows a Gaussian normal density function with mean $\mathbf{M}(G)^{-1}\mathbf{X}\boldsymbol{\delta}$ and variance $\mathbf{M}(G)^{-1}$.

[23] The existence of a unique potential maximizer $\mathbf{q}^*$ is guaranteed when the output levels are bounded and the matrix $\mathbf{M}(G)$ is positive definite [Byong-Hun, 1983; König et al., 2018].

G.5. Matrix Perturbation

For the computational implementation of the MCMC algorithm we will have to evaluate the inverse of $\mathbf{M}(G)$ whenever a link has been added or removed in the network G. To do this in an efficient way, the following lemma will be helpful.

Lemma 6. *Let $\mathbf{e}_i$ and $\mathbf{e}_j$ be the i-th and j-th unit basis vectors in $\mathbb{R}^n$. The matrix $\mathbf{e}_i\mathbf{e}_j^\top$ has a one in the ij-th position and zeros everywhere else. Define the matrix*

$$\mathbf{B}_{ij} \equiv \frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i}.$$

Adding a perturbation α to the matrix $\mathbf{A}$ in the ij-th and the ji-th position can be written as $\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\rho\mathbf{e}_j\mathbf{e}_i^\top$.

(i) The inverse of the perturbed matrix can be written as

$$(\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\alpha\mathbf{e}_j\mathbf{e}_i^\top)^{-1} = \mathbf{A}^{-1}-\alpha\mathbf{B}_{ij}-\alpha\frac{\left(\mathbf{A}^{-1}-\alpha\mathbf{B}_{ij}\right)\mathbf{e}_j\mathbf{e}_i^\top\left(\mathbf{A}^{-1}-\alpha\mathbf{B}_{ij}\right)}{1+\alpha\mathbf{e}_i^\top\left(\mathbf{A}^{-1}-\alpha\mathbf{B}_{ij}\right)\mathbf{e}_j}. \tag{G.33}$$

(ii) The determinant of the perturbed matrix can be written as

$$\det(\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\alpha\mathbf{e}_j\mathbf{e}_i^\top) = \left(1+\alpha\mathbf{e}_i^\top\left(\mathbf{A}^{-1}-\alpha\mathbf{B}_{ij}\right)\mathbf{e}_j\right)\left(1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i\right)\det\left(\mathbf{A}\right). \tag{G.34}$$

Proof of Lemma 6. We first give a proof of part (i) of the lemma. The Sherman–Morrison formula states that [Meyer, 2000, page 124]

$$(\mathbf{A}+\alpha\mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1}-\alpha\frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}.$$

Let $\mathbf{c}=\mathbf{e}_i$ and $\mathbf{d}=\mathbf{e}_j$, where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the i-th and j-th unit basis vectors, respectively. The matrix $\mathbf{c}\mathbf{d}^\top$ then has a one in the (i,j)-position and zeros elsewhere, so that adding a one to the matrix $\mathbf{A}$ in the (i,j)-position and the (j,i)-position (resulting from adding a link ij to the adjacency matrix $\mathbf{A}$) yields a perturbed matrix $\mathbf{B}$ which can be written as

$$\mathbf{B} = \mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\alpha\mathbf{e}_j\mathbf{e}_i^\top = \mathbf{C}+\alpha\mathbf{e}_j\mathbf{e}_i^\top,$$

where we have denoted by $\mathbf{C}\equiv\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top$. Using the Sherman–Morrison formula we then can write

$$\mathbf{B}^{-1} = (\mathbf{C}+\alpha\mathbf{e}_j\mathbf{e}_i^\top)^{-1} = \mathbf{C}^{-1}-\alpha\frac{\mathbf{C}^{-1}\mathbf{e}_j\mathbf{e}_i^\top\mathbf{C}^{-1}}{1+\alpha\mathbf{e}_i^\top\mathbf{C}^{-1}\mathbf{e}_j}, \tag{G.35}$$

while applying Sherman–Morrison to $\mathbf{C}^{-1}$ gives

$$\mathbf{C}^{-1} = (\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top)^{-1} = \mathbf{A}^{-1}-\alpha\frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i}.$$

Inserting $\mathbf{C}^{-1}$ into Equation (G.35) yields

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} - \alpha \frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i} - \alpha \frac{\left(\mathbf{A}^{-1} - \alpha\frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i}\right)\mathbf{e}_j\mathbf{e}_i^\top\left(\mathbf{A}^{-1} - \alpha\frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i}\right)}{1+\alpha\mathbf{e}_i^\top\left(\mathbf{A}^{-1} - \alpha\frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i}\right)\mathbf{e}_j}.$$

We next give a proof of part (ii) of the lemma. We have that $\det(\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\alpha\mathbf{e}_j\mathbf{e}_i^\top) = \det(\mathbf{C}+\alpha\mathbf{e}_j\mathbf{e}_i^\top)$, where we have denoted by $\mathbf{C} \equiv \mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top$. The matrix determinant lemma states that [Horn and Johnson, 1990]

$$\det(\mathbf{A}+\mathbf{u}\mathbf{v}^\top) = (1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u})\det(\mathbf{A}).$$

It then follows that

$$\det(\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\alpha\mathbf{e}_j\mathbf{e}_i^\top) = \det(\mathbf{C}+\alpha\mathbf{e}_j\mathbf{e}_i^\top) = (1+\alpha\mathbf{e}_i^\top\mathbf{C}^{-1}\mathbf{e}_j)\det(\mathbf{C}),$$

Similarly, from the matrix determinant lemma, we have that

$$\det(\mathbf{C}) = \det\left(\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top\right) = (1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i)\det(\mathbf{A}).$$

Further, the Sherman–Morrison formula states that [Meyer, 2000, page 124]

$$(\mathbf{A}+\alpha\mathbf{c}\mathbf{d}^\top)^{-1} = \mathbf{A}^{-1} - \alpha\frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}.$$

It follows that

$$\mathbf{C}^{-1} = (\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top)^{-1} = \mathbf{A}^{-1} - \alpha\frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i}.$$

Putting the above results together, we find that

$$\begin{aligned}\det(\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\alpha\mathbf{e}_j\mathbf{e}_i^\top) &= (1+\alpha\mathbf{e}_i^\top\mathbf{C}^{-1}\mathbf{e}_j)\det(\mathbf{C}) \\ &= \left(1+\alpha\mathbf{e}_i^\top\left(\mathbf{A}^{-1} - \alpha\frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i}\right)\mathbf{e}_j\right)\left(1+\alpha\mathbf{e}_j^\top\mathbf{A}^{-1}\mathbf{e}_i\right)\det(\mathbf{A}).\end{aligned}$$

□

Note that due to Lemma 6 the determinant and the inverse of the perturbed matrix $\mathbf{A}+\alpha\mathbf{e}_i\mathbf{e}_j^\top+\alpha\mathbf{e}_j\mathbf{e}_i^\top$ can be efficiently computed if the determinant, $\det(\mathbf{A})$, and the inverse, $\mathbf{A}^{-1}$, of $\mathbf{A}$ are already known.

H. Consistency, Computation Time and Missing Data

In this paper we consider three types of Bayesian MCMC estimation methods, the likelihood partition (LP), the double Metropolis-Hastings (DMH), and the adaptive exchange (AEX) algorithm to estimate our model. In order to demonstrate the performance of each method, we conduct a small-scale simulation study to show that (i) the true parameter values can be obtained from each method when estimating the correct model, and (ii) to illustrate the relative computation cost of each method. For the first part, as known from the Bayesian identification literature [see, e.g., Berger, 1985, p.224], the Bayesian asymptotic theory related to the posterior distribution is identical to the asymptotic distribution theory for the maximum likelihood estimator (MLE). Therefore, the posterior distribution of parameters should be concentrated at the true value, i.e.,

consistent, when the sample size goes to infinity (where the role of the prior vanishes), just like MLE does. Regarding the latter, since estimating ERGMs is usually computationally costly, we report the computation time (under our simulation design) so that potential users can better gauge the feasibility of our methods when applied to their own data.

In this simulation, we generate an artificial network (G) and output data ($\mathbf{q}$) from the data generating process (DGP) described in Definition 1. We consider two network sizes (n) – 100 and 200 – for comparing posterior distributions under different sample sizes. In the DGP, we capture individual exogenous heterogeneity in the output process by βx, where the variable x is generated from a normal distribution, $\mathcal{N}(0, 4)$, and the coefficient β is set to one. The true value of the complementarity effect ρ is set to 0.05 and the true value of the substitution effect b is set to 0.01. Exogenous linking costs for each network link (a_{ij}) are captured by a constant term and two dyad-specific exogenous variables c_1 and c_2 generated as follows: For c_1, we first draw n discrete uniform variables H from $[1, 4]$, If H_{1i} and H_{1j} are the same, then we set $c_{1,ij} = 1$. Otherwise, we set it to zero. For c_2, we use the exogenous variable x from the output process and define $c_{2,ij} = |x_i - x_j|$. The parameters assigned to the cost function $\gamma_0 + \gamma_1 c_{1,ij} + \gamma_2 c_{2,ij}$ are set to $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2) = (7, -2, -1)$. We normalize other parameters in Definition 1 as follows: $\vartheta = 1$, $\nu = 0.5$, $\lambda = 1$, $\xi = 1$, $\chi = 1$.

Given the artificial data, in terms of the output levels $\mathbf{q}$ and the network G, we first estimate the exact same model in the DGP and simulate $50,000$ MCMC draws from the conditional posterior distribution $P(\rho, b, \boldsymbol{\gamma}', \beta|\mathbf{q}, G)$ under the LP, DMH, and AEX methods discussed in Section 3.3 of the paper. We initially code all of the three methods in Matlab. However, we found Matlab was not ideal for handling the heavy sequential computations that are part of the DMH and AEX algorithms. Therefore, we recode the DMH and AEX methods in Fortran. The average computation time for a single MCMC iteration (measured in seconds), which is executed on a single workstation with dual Intel Xeon 2.60 GHz CPUs, is reported in Figure H.1. One can see that when the network size increases, the computation time required by the DMH and AEX algorithms increases dramatically compared to the LP algorithm, even though we have adopted a high level programming language such as Fortran in order to gain computational efficiency. Moreover, we illustrate the distributions of the posterior draws in Figure H.2. One can see that when the sample size increases from 100 to 200, the draws of (ρ, b, γ') become more concentrated around the true values.

Furthermore, we use this simulation environment to examine the impact of missing observations on the estimation results. We take the artificial network and output sample with size equal to 200 generated from the DGP described above and conduct 100 Monte Carlo repetitions. In each repetition, we remove different amounts of data, 25%, 50%, and 75% of the total observations, respectively, at random and estimate the model using the LP approach. We compute the posterior mean from MCMC draws as the point estimate for each parameter and report the mean and the standard deviation across the Monte Carlo repetitions in Table H.1. We find that the bias and volatility of the point estimates increase with increasing amounts of missing observations. However, from the most severe case (75% missing) that we consider, in which only 50 nodes were left in our sample, the average biases are maintained below 15% for the complementarity effect ρ; below 72% for the substitution effect b; and generally below 12% for the other parameters in the linking cost function. From this simulation exercise we therefore find that the bias incurred from missing data is relatively weak.

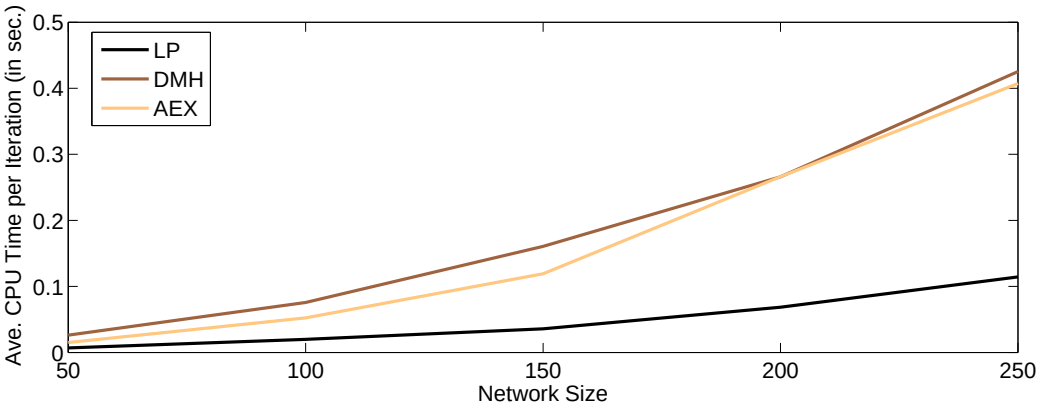


Figure H.1: The average computation time for a single MCMC iteration (measured in seconds).

Table H.1: Monte Carlo simulation results based on increasing levels of missing data.

	DGP	25% missing		50%missing		75% missing	
		mean	s.d.	mean	s.d.	mean	s.d.
ρ	0.0500	0.0578	0.0070	0.0677	0.0108	0.0414	0.0156
b	0.0100	0.0098	0.0020	0.0123	0.0033	0.0172	0.0040
γ_0	7.0000	7.0992	0.1374	7.1463	0.2371	7.4734	0.7513
γ_1	-2.0000	-1.9797	0.0807	-1.9851	0.1847	-2.0692	0.3760
γ_2	-1.0000	-1.0372	0.0373	-1.0450	0.0623	-1.1196	0.1894

Note: The number of repetitions for each simulation is set to 100. The true parameters are provided in the first column. We consider different levels of missing data: 25%, 50%, and 75%. The mean and the standard deviation of the point estimates across 100 repetitions are reported.

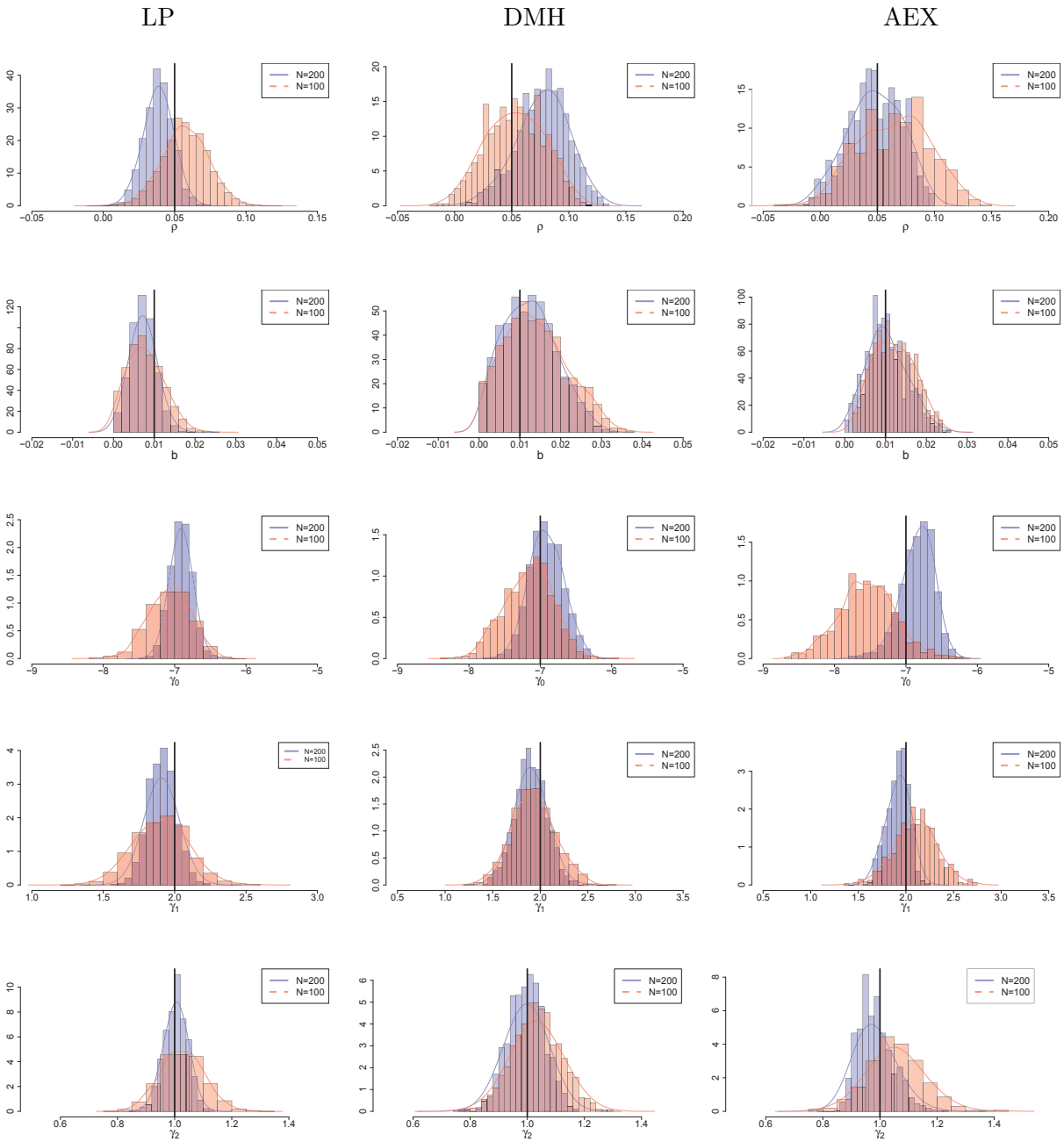


Figure H.2: The distribution of the parameters ρ, b, and γ across a total of $50,000$ MCMC draws from the conditional posterior distributions using the likelihood partition (LP) approach (left column), the double Metropolis-Hastings (DMH) algorithm (middle column) and the AEX algorithm (right column) discussed in Section 3.3.

References

Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly*, 45(3):425–455.

Anderson, S., Goeree, J., and Holt, C. (2004). Noisy directional learning and the logit equilibrium. *Scandinavian Journal of Economics*, 106(3):581–602.

Anderson, S. P., De Palma, A., and Thisse, J. F. (1992). *Discrete choice theory of product differentiation*. MIT Press.

Anderson, S. P., Goeree, J. K., and Holt, C. A. (2001). Minimum-effort coordination games: Stochastic potential and logit equilibrium. *Games and Economic Behavior*, 34(2):177–199.

Atkinson, K. E. (1989). *An introduction to numerical analysis*. John Wiley & Sons.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

Banerjee, A. and Duflo, E. (2005). Growth theory through the lens of development economics. *Handbook of Economic growth*, 1:473–552.

Baum, J., Cowan, R., and Jonard, N. (2009). Network-independent partner selection and the evolution of innovation networks. *UNU-MERIT working paper series; 2009-022*.

Belderbos, R., Carree, M., and Lokshin, B. (2004). Cooperative r&d and firm performance. *Research policy*, 33(10):1477–1492.

Ben-Akiva, M. and Watanatada, T. (1981). Application of a continuous spatial choice logit model. *Structural analysis of discrete data with econometric applications*, pages 320–343.

Bena, J., Fons-Rosen, C., and Ondko, P. (2008). Zephyr: Ownership changes database. *London School of Economics working paper*.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Berliant, M. and Fujita, M. (2008). Knowledge creation as a square dance on the hilbert cube. *International Economic Review*, 49(4):1251–1295.

Berliant, M. and Fujita, M. (2009). Dynamics of knowledge creation and transfer: The two person case. *International Journal of Economic Theory*, 5(2):155–179.

Bernstein, J. I. (1988). Costs of production, intra-and interindustry R&D spillovers: Canadian evidence. *Canadian Journal of Economics*, 21(2):324–347.

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.

Blundell, R., Griffith, R., and Van Reenen, J. (1995). Dynamic count data models of technological innovation. *The Economic Journal*, 105(429):333–344.

Bollaert, H. and Delanghe, M. (2015). Securities data company and zephyr, data sources for m&a research. *Journal of Corporate Finance*.

Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122.

Brualdi, R. A. and Solheid, Ernie, S. (1986). On the spectral radius of connected graphs. *Publications de l' Institute Mathématique*, 53:45–54.

Byong-Hun, A. (1983). Iterative methods for linear complementarity problems with upper bounds on primary variables. *Mathematical Programming*, 26(3):295–315.

Chatterjee, S., Diaconis, P., and Sly, A. (2011). Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21(4):1400–1435.

Chib, S. and Greenberg, E. (1996). Markov chain Monte Carlo simulation methods in econometrics. *Econometric theory*, 12(03):409–431.

Cohen, W. and Klepper, S. (1996a). Firm size and the nature of innovation within industries: the case of process and product R&D. *The Review of Economics and Statistics*, 78(2):232–243.

Cohen, W. and Klepper, S. (1996b). A reprise of size and R&D. *The Economic Journal*, 106(437):925–951.

Copeland, A. and Fixler, D. (2012). Measuring the price of research and development output. *Review of Income and Wealth*, 58(1):166–182.

Cvetkovic, D. and Rowlinson, P. (1990). The largest eigenvalue of a graph: A survey. *Linear and Multinilear Algebra*, 28(1):3–33.

Dai, R. (2012). International accounting databases on WRDS: Comparative analysis. *Working paper, Wharton Research Data Services, University of Pennsylvania*.

D'Aspremont, C. and Jacquemin, A. (1988). Cooperative and noncooperative R&D in duopoly with spillovers. *The American Economic Review*, 78(5):1133–1137.

Deijfen, M. and Kets, W. (2009). Random intersection graphs with tunable degree distribution and clustering. *Probability in the Engineering and Informational Sciences*, 23(04):661–674.

Dell, M. (2009). GIS analysis for applied economists. *Mimeo, MIT Department of Economics*.

D'Haultfœuille, X. and Iaria, A. (2016). A convenient method for the estimation of the multinomial logit model with fixed effects. *Economics Letters*, 141:77–79.

Dorogovtsev, S. N. and Mendes, J. F. (2013). *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press.

Fafchamps, M. and Gubert, F. (2007). The formation of risk sharing networks. *Journal of Development Economics*, 83(2):326–350.

Fort, G., Moulines, E., and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *The Annals of Statistics*, 36(4):3262–3289.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, pages 169–193.

Graham, B. S. (2014). An empirical model of network formation: detecting homophily when agents are heterogenous. *National Bureau of Economic Research Working Paper No. w20341*.

Graham, B. S. (2015). Methods of identification in social networks. *Annual Review of Economics*, 7(1):465–485.

Griffith, R., Redding, S., and Van Reenen, J. (2003). R&D and absorptive capacity: Theory and empirical evidence.

Scandinavian Journal of Economics, 105(1):99–118.
Hagedoorn, J. (2002). Inter-firm R&D partnerships: an overview of major trends and patterns since 1960. *Research Policy*, 31(4):477–492.
Hanaki, N., Nakajima, R., and Ogura, Y. (2010). The dynamics of R&D network in the IT industry. *Research Policy*, 39(3):386–399.
Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press.
Huyghebaert, N. and Luypaert, M. (2010). Antecedents of growth through mergers and acquisitions: Empirical results from belgium. *Journal of Business Research*, 63(4):392–403.
Jaffe, A. and Trajtenberg, M. (2002). *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press.
Jaffe, A. B. (1989). Characterizing the technological position of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, 18(2):87 – 97.
Jin, I. H., Yuan, Y., and Liang, F. (2013). Bayesian analysis for exponential random graph models using the adaptive exchange sampler. *Statistics and its interface*, 6(4):559.
Kamien, M. I., Muller, E., and Zang, I. (1992). Research joint ventures and R&D cartels. *The American Economic Review*, 82(5):1293–1306.
König, M., Tessone, C., and Zenou, Y. (2014a). Nestedness in networks: A theoretical model and some applications. *Theoretical Economics*, 9:695–752.
König, M. D., Liu, X., and Zenou, Y. (2014b). R&D networks: Theory, empirics and policy implications. *CEPR Discussion Paper No. 9872*.
König, M. D., Lorenz, J., and Zilibotti, F. (2016). Innovation vs. imitation and the evolution of productivity distributions. *Theoretical Economics*, 11:1053–1102.
Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian econometric methods*. Cambridge University Press.
Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2015). An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. *Journal of the American Statistical Association*, 111(513):377–393.
Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.
Lychagin, S., Pinkse, J., Slade, M. E., and Van Reenen, J. (2016). Spillovers in space: does geography matter? *The Journal of Industrial Economics*, 64(2):295–335.
Mahadev, N. and Peled, U. (1995). *Threshold Graphs and Related Topics*. North Holland.
McFadden, D. (1976). The mathematical theory of demand models. *Behavioral Travel Demand Models*, pages 305–314.
McManus, O., Blatz, A., and Magleby, K. (1987). Sampling, log binning, fitting, and plotting durations of open and shut intervals from single channels and the effects of noise. *Pflügers Archiv*, 410(4-5):530–553.
Mele, A. (2017). A structural model of dense network formation. *Econometrica*, 85(3):825–850.
Melitz, M. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725.
Melitz, M., Ottaviano, G., and Maggiore, S. (2008). Market Size, Trade, and Productivity. *Review of Economic Studies*, 75(1):295–316.
Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics.
Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., and Van Den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy*, 36(7):1016–1034.
Papadopoulos, A. (2012). Sources of data for international business research: Availabilities and implications for researchers. In *Academy of Management Proceedings*, volume 2012, pages 1–1. Academy of Management.
Peled, U. N., Petreschi, R., and Sterbini, A. (1999). (n, e)-graphs with maximum sum of squares of degrees. *Journal of Graph Theory*, 31(4):283–295.
Powell, W., Koput, K. W., and Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, 41(1):116–145.
Powell, W. W., White, D. R., Koput, K. W., and Owen-Smith, J. (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology*, 110(4):1132–1205.
Raftery, A. E., Lewis, S., Banfield, J. D., and Rafterv, A. E. (1992). How many iterations in the Gibbs sampler. *Bayesian Statistics*, page 4.
Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.
Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of applied probability*, 44(2):458–475.
Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.
Roijakkers, N. and Hagedoorn, J. (2006). Inter-firm R&D partnering in pharmaceutical biotechnology since 1975: Trends, patterns, and networks. *Research Policy*, 35(3):431–446.
Schilling, M. (2009). Understanding the alliance data. *Strategic Management Journal*, 30(3):233–260.
Singh, N. and Vives, X. (1984). Price and quantity competition in a differentiated duopoly. *The RAND Journal of Economics*, 15(4):546–554.
Snijders, T. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.
Söderberg, B. (2002). General formalism for inhomogeneous random graphs. *Physical Review E*, 66(6):066121.
Spence, M. (1984). Cost reduction, competition, and industry performance. *Econometrica*, 52(1):101–122.
Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B. H., and Harhoff, D. (2010). Harmonizing and combining large datasets-an application to firm-level patent and accounting data. *National Bureau of Economic Research Working Paper No. w15851*.
Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 22(4):1701–1728.
Trajtenberg, M., Shiff, G., and Melamed, R. (2009). The "names game": Harnessing inventors, patent data for economic research. *Annals of Economics and Statistics*, (93/94):79–108.
Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations.

Survey review, 23(176):88–93.
Westbrock, B. (2010). Natural concentration in industrial research collaboration. *The RAND Journal of Economics*, 41(2):351–371.
Wong, R. (2001). *Asymptotic approximations of integrals*, volume 34. Society for Industrial Mathematics.